ISSI 2019

SAPIENZA
UNIVERSITÀ DI ROMA

17th INTERNATIONAL CONFERENCE ON SCIENTOMETRICS & INFORMETRICS

ISSI2019

with a Special STI Indicators Conference Track

2-5 September 2019
Sapienza University of Rome, Italy

PROCEEDINGS

VOLUME I

EdizioniEfesto

PROCEEDINGS OF THE 17TH CONFERENCE OF THE INTERNATIONAL SOCIETY FOR SCIENTOMETRICS AND INFORMETRICS

© Edizioni Efesto - ISBN 978-88-3381-118-5 - August 2019
Printed in Italy

Editors: *Giuseppe Catalano, Cinzia Daraio, Martina Gregori, Henk F. Moed and Giancarlo Ruocco*

Graphic cover design: *Francesco Manzo | graframan.com*
Cover photo: ©*Fayee - stock.adobe.com*

INTRODUCTION TO THE ISSI2019 CONFERENCE PROCEEDINGS

The 17th International Conference of the International Society for Scientometrics and Informetrics (ISSI2019) is held on 2-5 September 2019 in Rome, and is hosted by the Sapienza University of Rome. It is a major event with participants from 44 countries from all global regions.

The conference includes a special STI Indicators Conference Track organized in collaboration with the European Network of Indicator Designers (ENID). In this way, ISSI2019 represents a first experiment to bring together the two conferences in a particular year.

In a first round, around 420 submissions were made of full papers and research-in-progress papers. After an extensive peer review process, thoroughly conducted by over 200 reviewers, some 260 submissions were accepted for oral presentation, while the authors of another 80 submissions were invited to present their paper as a poster. Of these 80, about 60 per cent accepted this invitation. A second round of poster submissions was held, resulting in some 130 new poster submissions.

In the final acceptance decision, apart from the reviewer judgments, two additional rules were applied: for oral presentations of papers a one-presentation-per-presenter rule, and for papers and posters the rule that all contributions must be included in the conference proceedings. All in all, the Conference Proceedings contain 261 oral presentations of accepted papers, and 156 poster presentations.

The Conference Proceedings consist of four parts:

» Table of contents
» Keynotes
» Papers (full papers and research-in-progress)
» Posters
» Author Index

Papers and posters are ordered by their submission number in Easychair. The Author Index includes both the first author and co-authors of each indexed contribution. For each name, a list is given of the initial page numbers (in the conference proceedings) of the presented contributions.

Although the overwhelming part of the papers and posters comply with the format specified in advance on the conference website, there are several cases in which the final version sent by the authors does not have this prescribed format. To maximize the information content and usefulness of our conference proceedings, we decided to include such versions with deviant formats as well.

We first of all wish to thank the Magnifico Rettore of the Sapienza University of Rome, Prof. Eugenio Gaudio, and his university staff to make the university infrastructure available for the organization of this conference. Next, we thank all authors for their submissions, and the members of the Scientific Committee for their efforts in the peer review process. Special thanks go to all Committees and Program Committee members. In particular, we acknowledge Cassidy Sugimoto for the organization of the Doctoral Forum, Jacqueline Leta for the organization of the Poster sessions, Ed Noyons for his contribution to the organization of the Special STI-ENID Track, and Kevin Boyack for his collaboration to the Workshops and Tutorials organization. We wish to thank our Sponsors for their support that made an important contribution to the organization of this conference. Last but not least, we are grateful to Martina Gregori for her enormous efforts in the technical editing process of these Conference Proceedings and to Riccardo Cervelli for his contribution to the creation of the Author Index.

On behalf of the organizers and co-program Chairs of ISSI2019, Giuseppe Catalano, Cinzia Daraio and Giancarlo Ruocco,

Henk F. Moed
Program Chair ISSI2019

ORGANIZING COMMITTEE

Giuseppe Catalano, Sapienza University of Rome (Italy)
Cinzia Daraio, Sapienza University of Rome (Italy)
Giancarlo Ruocco, Sapienza University of Rome (Italy)
Henk F. Moed, Sapienza University of Rome (Italy)

PROGRAM COMMITTEE: CHAIRS

Henk F. Moed, Sapienza University of Rome (Italy), *Program Chair*
Giuseppe Catalano, Sapienza University of Rome (Italy), *Program co-Chair*
Cinzia Daraio, Sapienza University of Rome (Italy), *Program co-Chair*
Giancarlo Ruocco, Sapienza University of Rome (Italy), *Program co-Chair*

PROGRAM COMMITTEE: MEMBERS

Kevin Boyack, SciTech Strategies Inc. (USA), *Chair of Workshops and Tutorials*
Guillaume Cabanac, University of Toulouse (France), *Doctoral Forum*
Nicolas Carayol, University of Bordeaux (France)
Pei-Shan Chi, Katholieke Universiteit Leuven (Belgium), *Poster Session*
Jordi Molas Gallart, Universitat Politècnica de València (Spain)
Wolfgang Glänzel, Katholieke Universiteit Leuven (Belgium)
Sybille Hinze, German Centre for Higher Education Research and Science Studies (Germany)
Vincent Larivière, Université de Montréal (Canada), *Doctoral Forum*
Jacqueline Leta, Federal University of Rio de Janeiro (Brazil), *Chair Poster Session*
Ed Noyons, Leiden University (Netherlands), *Chair Special STI Track*
Ronald Rousseau, Katholieke Universiteit Leuven (Belgium)
Andrea Scharnhorst, Royal Netherlands Academy of Arts and Sciences (Netherlands), *Doctoral Forum*
Cassidy Sugimoto, Indiana University (USA), *Chair Doctoral Forum*
Lin Zhang, Wuhan University (China)
Alesia Zuccala, University of Copenhagen (Denmark), *Doctoral Forum*

EUGENE GARFIELD DOCTORAL DISSERTATION AWARD 2019 COMMITTEE

Birger Larsen, University of Aalborg, Denmark (ISSI Board member; Chair)
Nees Jan van Eck, CWTS, Leiden University, the Netherlands (ISSI Board member)
Elías Sanz Casado, Carlos III University of Madrid, Spain
Gemma Derrick, Lancaster University, United Kingdom

COMMITTEE FOR THE ISSI STUDENT TRAVEL AWARD

Aparna Basu, South Asian University, India (ISSI Board member)
Kevin Boyack, SciTech Strategies, Inc., USA (ISSI Board member, Chair)
Grant Lewison, Kings College London, UK (ISSI Board member)

SCIENTIFIC COMMITTEE

Kaare **Aagard**
Giovanni **Abramo**
Jonathan **Adams**
Euan **Adie**
Isidro **Aguillo**
Per **Ahlgren**
Isola **Ajiferuke**
Aliakbar **Akbaritabar**
Dag **Aksnes**
Juan-Pablo **Alperín**
Valeria **Aman**
Jens-Peter **Andersen**
Clément **Arsenault**
Rigas **Arvanitis**
Fredrik **Astrom**
Joaquin **Azagra-Caro**
Luiza **Badin**
Tomas **Baiget**
Rafael **Ball**
Omar **Ballester-Gonzalez**
Judit **Bar-Ilan**
Tomaz **Bartol**
Aparna **Basu**
Federico **Bianchi**
Johan **Bollen**
Andrea **Bonaccorsi**
Maria **Bordons**
Lutz **Bornmann**
Nelius **Boshoff**
Timothy **Bowman**
Kevin **Boyack**
Barry **Bozeman**
Susanne **Buehrer**
Clara **Calero-Medina**
Julie **Callaert**
Juan-Miguel **Campanario**
David **Campbell**
Carolina **Cañibano**
Nicolas **Carayol**
Camila **Carneiro-Dias**
Niccolo **Casnici**
Giuseppe **Catalano**
Federico **Caviggioli**
Ching-Chun **Chang**
Diego **Chavarro**
Daniele **Checchi**
Chaomei **Chen**
Dar-Zen **Chen**
Pei-Shan **Chi**
Zaida **Chincilla-Rodriguez**
Tommaso **Ciarli**
Giovanni **Colavizza**
Hugo **Confraria**
Rodrigo **Costas**
Grégoire **Côté**
Regino **Criado**
Andrea **D'Angelo**
Hans-Dieter **Daniel**
Cinzia **Daraio**
Koenraad **Debackere**
Nicola **De-Bellis**
Felix **De Moya**
Tahereh **Dehdarirad**
Gaetan **De-Rassenfosse**
Gemma **Derrick**
Carole **Des-Bordes**
Adrián-Arias **Díaz-Faes**
Ying **Ding**
Paul **Donner**
Ekaterina **Dyachenko**
Tim **Engels**
Adian **Fatchur-Rochim**
Márcia **Ferreira-Goncalves**
Thomas **Franssen**
Chiara **Franzoni**
Rainer **Frietsch**
Stephan **Gauch**
Daphne **Getz**
Aldo **Geuna**
Gita **Ghiasi**
Elea **Giménez-Toledo**
Wolfgang **Glänzel**
Jochen **Gläser**
Abdullah **Gok**
Alicia **Gomez**
Juan **Gorraiz**
Martina **Gregori**
Miguel **Guevara**
Raf **Guns**
Gali **Halevi**
Björn **Hammarfelt**
Robin **Haunschild**
Stefanie **Haustein**
Gaston **Heimeriks**
Diana **Hicks**
Sybille **Hinze**
Kim **Holmberg**
Xiaojun **Hu**
Mu-Hsuan **Huang**
Masatsura **Igami**
Margriet **Jansz**
Wei **Jeng**
Evaristo **Jimenez-Contreras**
Ben **Jongbloed**
Koen **Jonkers**
Wolfgang **Kaltenbrunner**
Stijn **Kelchtermans**
Dick **Klavans**
Stacy **Konkiel**
Kayvan **Kousha**
Emanuel **Kulczycki**
Wout **Lamers**
Philippe **Laredo**
Vincent **Larivière**
Birger **Larsen**
Grit **Laudel**
Cornelia **Lawson**
Steffen **Lemke**
Maurizio **Lenzerini**
Benedetto **Lepori**
Jacqueline **Leta**
Loet **Leydesdorff**
Judith **Licea**
Rodrigo **Liscovsky-Barrera**
Francesco **Lissoni**
Patrick **Llerena**
Carmen **Lopez-Illescas**
Marc **Luwel**
Domenico **Maisano**
Alberto **Martin-Martin**
Camilla **Mastromarco**
Diana **Maynard**
Athanasios **Mazarakis**
Julia **Melkers**
Michele **Meoli**
Ernest **Miguelez**
Stasa **Milojevic**
Egidio **Miotti**
Henk **Moed**
Jordi **Molas Gallart**

Philippe **Mongeon**
Fernanda **Morillo**
Olga **Moskaleba**
Johann **Mouton**
Alexis-Michel **Mugabushaka**
Rogerio **Mugnaini**
Ulle **Must**
Rudiger **Mutz**
Tina **Nane**
Anton **Nederhof**
Ed **Noyons**
Kaltrina **Nuredini**
Michael **Ochsner**
Luis **Orozco**
José-Luis **Ortega**
Astrid **Orth**
Stefano **Paleari**
Andres **Pandiella**
Laurens **Patricia**
Gabriele **Pellegrino**
Orion **Penner**
Antonio **Perianes-Rodríguez**
Bluma **Peritz**
Viola **Peter**
Isabella **Peters**
Diogo **Pinheiro**
Vladimir **Pislyakov**
Guilherme-Ary **Plonski**
Heidi **Prozesky**
Francesco **Quatraro**
Julio **Raffo**
Ismael **Rafols**
Emanuela **Reale**
Nicolas **Robinson-Garcia**
Matteo **Romanello**
Daniele **Rotolo**
Ronald **Rousseau**
John **Ruggiero**
Giancarlo **Ruocco**
Alex **Rusforth**
Frederique **Sachwald**
Maxime **Saint-Marie**
Ulf **Sandström**
Shivappa **Sangam**
Elias **Sanz**
Andrea **Scharnhorst**
Ad **Scheepers**
Thomas **Scherngell**
Edgar **Schiebel**
Christian **Schloegl**
Marion **Schmidt**
Ulrich **Schmoch**
Jesper **Schneider**
András **Schubert**
Torben **Schubert**
Fei **Shu**
Elena **Šimukovic**
Gunnar **Sivertsen**
Stig **Slipersæter**
Nynke-Jo **Smit**
Flaminio **Squazzoni**
Valerio **Sterzi**
Cassidy **Sugimoto**
Li **Tang**
Jakob **Tesch**
Henk **Test**
Mike **Thelwall**
Bart **Thijs**
Robert **Tijssen**
Daniel **Torres-Salinas**
Vincent **Traag**
Barbara **van-Balen**
Peter **van-Den-Besselaar**
Barend **van-der-Meulen**
Nees-Jan **van-Eck**
Thed **van-Leeuwen**
Bart **van-Looy**
Anthony **van-Raan**
Benjamin **Vargas-Quesada**
Liwen **Vaughan**
Theresa **Velden**
Niki **Vermeulen**
Antonio **Vezzani**
Peter **Vinkler**
Martijn **Visser**
Cathelijn **Waaijer**
Caroline **Wagner**
Ludo **Waltman**
Lili **Wang**
Xianwen **Wang**
Eric **Welch**
Jos **Winnink**
Richard **Woolley**
Angela **Wroblewski**
Erjia **Yan**
Alfredo **Yegros**
Jan **Youtie**
Zohreh **Zahedi**
Lin **Zhang**
Dangzhi **Zhao**
Ping **Zhou**
Lorenzo **Zirulia**
Alesia **Zuccala**

INDEX OF PAPERS (FULL PAPERS AND RESEARCH IN PROGRESS)

INDEX OF POSTERS

KEYNOTE LECTURES

KEYNOTE LECTURE 1

Quantification – Its Affordances and Limits

John Carson, University of Michigan (USA)

Abstract

We live in a world awash in numbers. Tables, graphs, charts, Fitbit readouts, spreadsheets that overflow our screens no matter how large, economic forecasts, climate modeling, weather predictions, journal impact factors, H-indices, and the list could go on and on, still barely scratching the surface. We are measured, surveyed, and subject to constant surveillance, largely through the quantification of a dizzying array of features of ourselves and the world around us. In this talk I will draw on work I have done on the quantification and measurement of intelligence to discuss some of the insights that these processes of quantification can bring, but also some of the perils that inevitably accompany them. The story of how intelligence became a measurable quality and other examples from the history of quantification suggest that quantification and measurement should be seen not just as technical pursuits, but also as normative ones. Every act of seeing, whether through sight or numbers, is also an act of occlusion, of not-seeing. And every move to make decisions more orderly and rational by translating a question into numerical comparisons is also a move to render irrelevant and often invisible the factors that were not included. The reductions and simplifications quantifications rely on can without question bring great and important clarity, but always at a cost. Among the moral questions for the practitioner, I will suggest, is not just whether that cost is justified, but, even more critically, who is being asked to pay it? Whose sight is valued, and what picture of the world results from foregrounding those forms of quantification as opposed to some others, or perhaps none at all?

Biographical Sketch — John Carson

John Carson is Associate Professor of History at the University of Michigan, where he has been since 1998. He received his PhD in History (of Science) from Princeton University in 1994, and has held postdoctoral fellowships from the Wellcome Institute for the History of Medicine, Department of Science & Technology Studies at Cornell University, National Humanities Center, Wellesley College Newhouse Center for the Humanities, Max Planck Institute for the History of Science, and Wissenschaftskolleg, where he was part of a research group on the history of quantification. His current research project explores the development and deployment of the medico-legal category "unsoundness of mind" (*non compos mentis*) in the eighteenth and nineteenth centuries in England and America. More broadly he is interested in the history of the human sciences, quantification, and the production of norms. His publications include *The Measure of Merit: Talents, Intelligence, and Inequality in the French and American Republics, 1750-1940* (2007); "Mental Testing in the Early Twentieth Century: Internationalizing the Mental Testing Story," *History of Psychology* 17 (2014); and "Every Expression Is Watched: Mind, Medical Expertise and Display in the Nineteenth-Century English Courtroom," *Social Studies of Science* 48 (2018).

KEYNOTE LECTURE 2

New developments in scientometric and informetric research

Mike Thelwall, University of Wolverhampton (UK)

Abstract

This talk will discuss three recent developments that are changing scientometric practice: altmetrics, full text mining, and the impact agenda. After decades of research into altmetrics and webometrics, alternative indicators have emerged as a standard part of scholarly communication infrastructure. This can be seen in the availability of Altmetric.com scores in many publisher websites and the informal use of altmetrics supporting research evaluation narratives. Altmetrics are well enough understood that we can now recommend appropriate uses, and are ready to fully exploit them. Full text mining is a second development that is taking advantage of the increasing availability of collections of open access documents, such as the PubMed Central Open Access Subset, to get fine-grained information about citation contexts. This approach has the potential to identify important types of citation to get more precise evidence of the type of impact reflected by citation counts. The impact agenda in the UK Research Excellence Framework (REF) is the requirement for researchers to produce evidence-based narratives describing how their research has non-academic societal benefits. This has led to changes within UK universities to promote societal impact, such as through the appointment of impact officers. Altmetric and Webometric indicators are used as part of the evidence base of some of these narratives, which provide a tough challenge for scientometricians. The talk will finish with a few words about the importance of using research indicators responsibly. Practitioners should always be aware of potential biasing and systemic effects that can lead to unintended consequences.

Biographical Sketch — Mike Thelwall

Mike Thelwall, Professor of Data Science at the University of Wolverhampton in the UK, is the author of the cheap but excellent book *Web Indicators for Research Evaluation: A Practical Guide*. This book describes how to analyse altmetric and webometric data for research evaluation purposes, including appropriate methods for field normalising it and reporting it for use in evaluations. It is designed for scientometric researchers, practitioners and Master's level students. Mike is also the programmer of the free but excellent scientometric software *Webometric Analyst* http://lexiurl.wlv.ac.uk, which includes hundreds of functions to gather and process scientometric, altmetric and webometric data. For example, if you already have a set of bibliometric records to evaluate, Webometric Analyst can add scores from Altmetric.com, reader counts from Mendeley, citation counts from Google Books, or syllabus citations from the web. With Webometric Analyst you can then summarise this data in various ways, including with tables of field normalised indicators and confidence intervals. Webometric Analyst is a bit confusing to use at first because of its range of capabilities but people that both study the book and use the program can expect long, productive and fulfilling scientometric careers.

ISSI2019 PAPERS
(FULL PAPERS AND RESEARCH IN PROGRESS)
AND POSTERS

Towards Machine Readable Academic Biographies: A Deep Learning Approach

Patrick Kenekayoro[1]

[1] *patrick.kenekayoro@outlook.com*
Niger Delta University, Amassoma, PMB 581, Bayelsa State, Nigeria

Abstract

Free text biographies make up a percentage of the academic web. A number of studies have explored ways to extract semantic meaning from these free text academic biographies and have also cited the importance of this task. Some of these studies used natural language processing techniques to extract named entities, however it may also be beneficial to go a step further and represent the relationships between these named entities in a form that could be understood by computers. This will enable the execution of complex search or filter queries on academic biographies that are otherwise difficult to achieve in unstructured plain text academic biographies. The consensus is that the semantic web is the desired goal where web resources are represented in machine readable forms. Thus, this research demonstrates the use of deep learning techniques to extract academic biographies from personal web pages of researchers, identify the named entities in these academic biographies and then represent the relationships between the named entities in a semantic web standard that can be easily processed by computer algorithms.

Introduction

A number of studies have investigated ways to identify named entities in academic biographies (Kenekayoro, 2018; Tang et al., 2008), citing its importance for finding an alternate data source for Scientometrics/Informetrics researches, as the web is arguably the largest freely available data source. The study (Kenekayoro, 2018) used probabilistic machine learning techniques to classify word tokens that appeared in academic biographies in ORCID profiles into predetermined named entities. However, ORCID profiles were available in structured XML and JSON formats, which makes extracting biographies from these profiles a simpler task compared to extracting biographies from unstructured/semi structured texts on the web. An effective way to extract biographies from web pages is necessary before techniques such as those used in (Kenekayoro, 2018) can be used to identify named entities. This study can be seen as a follow up to (Kenekayoro, 2018), that addresses some of the questions posed in that study.

Kenekayoro (2018) highlighted the possibility of using the relationships between the named entities that appeared in academic biographies to give structure to plain text biographies, which in turn will make these biographies easier to process with computer algorithms. Nevertheless, these biographies sometimes need to be previously extracted from web pages, thus, this study aims to

1. Identify an effective way to extract academic biographies from the personal web pages of researchers.
2. Identify an effective way to create a structure of semantic connections between the named entities in academic biographies.

Related work

The task to automatically process researchers' web pages is identical to the preprocessing of Curriculum Vitae (CV) or researchers, thus methods used for these problems are transferable. Yu, Guan and Zhou (2005) have used support vector machines and Hidden Markov Models to parse and segment the contents of CVs into structured blocks with good results.

Kenekayoro (2018) identified the academic named entities that may appear in the biographies of researchers. Using support vector machines and random forests supervised learning

algorithms, the author extracted these named entities from plain text biographies. Supervised learning algorithms require the identification of fine grained features for optimal performance, however with the use of deep learning strategies, the algorithm may be able to identify necessary features required for optimal performance during the course of training the deep learning network. Relations between extracted named entities can then be represented using semantic web technologies to enable machine processing, which otherwise will be difficult to achieve in plain text academic biographies. Subsequent sections give an overview of deep learning and semantic web technologies.

Deep learning

Supervised learning techniques have been used in a number of Scientometrics and Webometrics researches (Han et al., 2003; Kenekayoro, Buckley, & Thelwall, 2014), particularly in the classification of academic web pages (Hernández, Rivero, Ruiz, & Corchuelo, 2016; Kenekayoro et al., 2014). A critical step in traditional machine learning algorithms is the identification of features that will be used to map an instance to a particular output. For example, Kenekayoro (2018) used Part of Speech (POS) tags and word endings as part of the features to identify the named entities in academic biographies. Identifying appropriate features requires domain knowledge of the problem to be solved, although there are feature selection techniques (Tang, Alelyani, & Liu, 2014) which could be used to find a subset of relevant features that can make supervised learning algorithms more accurate.

Deep learning, which is a subset of machine learning has being widely used in a number of scientific fields such as biomedicine (Mamoshina, Vieira, Putin, & Zhavoronkov, 2016) and traffic control systems (Fadlullah et al., 2017). Conceptually, a deep learning network is essentially a neural network (Zhang, 2000) with additional hidden layers, utilising the same principles such as backward propagation (Goh, 1995; Rumelhart, Hinton, & Williams, 1986) for training. Deep learning networks have an advantage over traditional machine learning algorithms in that, it is not necessary to explicitly define the features, or use feature selection strategies to identify the most relevant subset of features for the learning problem. Given a dataset, a deep learning network can automatically identify important features for the classification task. Yosinki et al. (2015) have also described a way to visualise hidden layers of neural networks for a better understanding of how neural networks identify relevant features.

In the application of deep learning approaches for Scientometrics or Webometrics research, techniques in the use of deep learning for natural language processing can be transferred, particularly for Webometrics research where the data is largely textual. Strategies such as word to vector (Pennington, Socher, & Manning, 2014), where words are represented as vectors with semantically similar words having identical vectors and recurrent neural networks (RNN) (Rumelhart et al., 1986; Werbos, 1988) that enables effective modelling of sequential data can be applied to improve the efficiency of webometrics research. In recurrent neural networks, the output of each time step in a sequence feeds the succeeding time step, thereby modelling how the input in various time steps depend on each other. For long sequences, it becomes difficult for recurrent neural networks to identify long term dependencies (Bengio, Simard, & Frasconi, 1994), so variations of RNN such as Long Short Term Memory (LSTM) (Schmidhuber & Hochreiter, 1997) and Gated Recurrent Units (GRU) (Cho, Van Merriënboer, Bahdanau, & Bengio, 2014) are used to effectively train long sequential data. Sequence to sequence modelling (Sutskever, Vinyals, & Le, 2014) is widely used for machine translation (Cho et al., 2014), however the same procedure can be used to solve the named entity recognition task in (Kenekayoro, 2018) where a sentence is translated to a sequence of Inside Outside Beginning (IOB) tags that represent the named entity of the corresponding index in an input sequence.

Deep learning techniques are applied in this study to extract biographies from academic web pages and the results are compared with traditional machine learning methods. A caveat is that training deep learning networks can be computationally expensive (Najafabadi et al., 2015), however development in hardware technologies and the availability of optimised deep learning frameworks such as Tensorflow (Abadi et al., 2016) reduces the hardware entry level for training deep learning networks.

Semantic web

The majority of web pages are written in Hyper Text Markup Language (HTML) that was originally designed for presenting documents in a human readable form, which is why it is difficult to extract information from HTML documents for machine processing. Automatically extracting information from HTML documents requires the use of web scraping techniques (Polidoro, Giannini, Conte, Mosca, & Rossetti, 2015) that may not be transferable across different web pages. The goal of the semantic web is to create a web where computers can efficiently extract and process web data. Several technologies have been developed to help move the web from its simple human readable form to a web where data is semantically connected, with the possibility for machines to efficiently manipulate this web data.

The Resource Description Framework (RDF); a W3C recommendation in 2004 enables storing relationships as triples in the form (Subject – Predicate – Object), where the predicate depicts the relationship between the subject and the object. An extension of the RDF is the Resource Description Framework Schema (RDFS) which adds expression to the basic RDF triples, there by creating the possibility for implementing reasoning through inference (Shadbolt, Berners-Lee, & Hall, 2006). The Web Ontology Language (OWL) adds more elaborate expressions for the relations between objects represented as classes. This creates the possibility for describing richer properties such as symmetry among objects (McGuinness, Van Harmelen, & others, 2004), thus enabling the possibility for determining if a particular concept is described in a knowledge base (ontology) or if the knowledge base is logically consistent (Shadbolt et al., 2006).

Knowledge representation using semantic web technologies has been used to represent e-science provenance in life science research. Provenance is invaluable as it describes the steps in an experiment that led to a result (Freire, Koop, Santos, & Silva, 2008), ensuring that the experiment can be repeated. Among other research areas, semantic web technologies have also been used in e-learning (Gladun et al., 2009) and business management (Hepp, Leymann, Domingue, Wahler, & Fensel, 2005). In this study, semantic web technologies are used to represent the relations between academic named entities in the biographies of researchers that were previously extracted from HTML web pages.

Methods

A homepage of an academic as shown in (Tang, Zhang, & Yao, 2007) is typically made up of the researcher's photo, contact information, a brief biography and a list of publications. The content of this page usually contains scripts and HTML formatting tags which can be regarded as noise because they are not useful in analysing the textual content of web pages. There are standard methods to strip these scripts and HTML tags from web pages so that the web pages contain only plain text information. However, as this study is interested in only the biographies in academic researchers' web pages, it is necessary to further exclude information such as publication lists, courses teaching etcetera.

Dataset

Tang et al., (2007) described the generation of a dataset available from (Tang et al., 2008), which annotated the words that appeared in the web pages of researchers into categories. The top level categories can be used to suggest the section sentences that appear in that block may belong to. Machine learning strategies can thus be used to identify patterns that represent respective sections in a researcher's personal web page, which can then be used to automatically group sentences from a typical academic researcher's web page into their respective sections. The dataset had a total of 23,358 sentences, with the majority of sentences part of the publication list of the researcher.

The top level sections identified in (Tang et al., 2008) are shown in Table 1. It is to be noted that on investigation of the dataset, the content of the introduction top level category showed to contain the biography of the researcher that owned the web page. Hence, the problem of extracting biographies from academic researchers' web pages is reduced to classifying sentences into the categories shown in Table 1 and retrieving the sentences that belong to the each top level category. To achieve this using machine learning strategies, it is necessary to:

- split the plain text data into individual sentences,
- and then identify those sentences that may be part of a biography.

Table 1.Top level categories extracted from an annotated dataset (Tang et al., 2008) of web pages of academics and the number of sentences belonging to the respective category

Top Level Categories	*Description*	*Size*
CONTACTINFO	Contains office address, phone number and email address.	1288
PUBLICATION	List of publications and/or attended conferences.	15229
RESINTERESTS	Research specialisation/disciplines.	1767
INTRODUCTION	Usually contains the biography of an academic researcher.	5074

Web page classification

To address the first aim of this study, which is identifying an effective way to extract academic biographies from web pages, two approaches; a deep learning network and traditional supervised machine learning algorithms were used to categorise sentences into a section that they may appear in an academic's web page. Both approaches are achieved using identical procedures.

Text extraction

The dataset (Tang et al., 2008) used in this study contains extracted and tagged texts from the web pages of 898 researchers. This tagged dataset can be used to create a machine learning model that will automatically tag preprocessed texts represented as a list of sentences, which have previously been extracted from web pages using functions such as Html2Text in Python or its equivalent in other high level programming languages. In this study, the Natural Language Processing Toolkit (Bird, Edward, & Ewan, 2009) is used to tokenize texts into sentences.

Feature extraction

Feature extraction is particularly important for traditional supervised learning algorithms, as the subset of features used greatly influence the accuracy of the learning model (Kursa & Rudnicki, 2010). On inspection of the dataset, it was noted that pronouns such as "I" and "my" appeared in sentences such as *"I worked on the Digital Michelangelo project"* that belonged to an academic biography, while sentences such as *"Investigation of Feature Selection Techniques"* that belonged to the publication section of a web page are usually made up of nouns and adjectives. The case or tense of words in a sentence can be additional attributes that may be useful features for supervised learning algorithms. The NLP toolkit (Bird et al., 2009) implements an algorithm that tags words in a sentence into the POS tags described in the PENN Treebank project (Marcus et al., 1994), which also incorporates the tense of words in its representation. Thus, for the traditional supervised learning techniques, an instance is represented as the number of each POS tag in the sentence. Deep learning networks do not necessarily need fine tuned features because relevant features can be identified during the course of training, thus instances for the deep learning algorithm are represented as a sequence of words, in the order in which they appeared in the input sentence.

Classification

Random Forests (RFs) (Breiman, 2001) and Support Vector Machines (SVMs) (Cortes & Vapnik, 1995), have been used in webometrics and bibliometric studies (Han et al., 2003; Kenekayoro et al., 2014), and also in author disambiguation (Culotta, Kanani, Hall, Wick, & McCallum, 2007; Han, Giles, Zha, Li, & Tsioutsiouliklis, 2004) and institution disambiguation (Cuxac, Lamirel, & Bonvallot, 2013) of scientific publications. SVMs on average, among the best performing machine learning techniques (Kenekayoro, 2018; Kenekayoro et al., 2014), even though it may be outperformed by other models in some problems (Caruana & Niculescu-Mizil, 2006). Thus, these supervised learning algorithms (SVMs and RFs), with their implementation in a python machine learning toolkit (scikit-learn) (Pedregosa et al., 2011) are used in this study.
The summary of the deep learning network implemented with Tensorflow (Abadi et al., 2016) that is used for classification is shown in Table 2.

Table 2. Summary of the deep learning network for the classification of sentences in a researchers' web page into its respective section

Layer (type)	*Output Shape*	*Number of Parameters*
Embedding	(None, 50, 50)	1,688,950
LSTM	(None, 30)	9720
Dense	(None, 4)	124

The first layer is an embedding layer that transforms a word in the sequence of words into a vector. The weights of the embedding layer is initialised using a pretrained glove vector of 50 dimensions (Pennington et al., 2014). The embedding layer ensures that similar words will be represented by identical vectors. Pennington et al. (2014) highlights that semantically similar words can be found by the cosine similarity of their word vectors. The output of the embedding layer is fed to an LSTM layer and then to a dense layer that outputs the final category a sentence belongs to by the softmax regression activation function. The results of classification with supervised learning and deep learning networks is reported in subsequent sections.

Results and Discussion

Table 3 and Table 4 show the accuracy of supervised learning models in the classification of the web page dataset. Two thirds of the dataset is used in training and a third in tests, with training accuracy determined by a 10 fold cross validation. Precision, recall and F-Score are additional metrics that show how well a learning model identifies individual classes in a dataset. A supervised learning model's precision for a category; A, is a metric that shows the extent to which the model correctly identifies instances that belong to Category A, while recall is the extent to which instances not belonging other categories are not wrongly classified as Category A.

SVMs have a number of parameters that could be tweaked to improve the accuracy of the resulting learning model. The results reported in Table 3 are for the optimal parameters identified using a brute force strategy (Grid Search) that compares the result of different parameter combinations to find the most appropriate for a learning problem.

Table 3. Accuracy of supervised and deep learning techniques for the extraction of biographies from academic web pages

Algorithm (Train, Test) accuracy	*Precision*	*Recall*	*F-Score*
Deep Learning (0.87, 0.88)	0.88	0.88	0.88
Random Forests (0.82, 0.82)	0.81	0.82	0.81
Support Vector Machines (0.80, 0.80)	0.81	0.80	0.80

The average precision, recall and f-score for all classes is reported in Table 3, but as the deep learning model outperformed the traditional supervised learning models, Table 4 shows the performance of the deep learning network in identifying individual categories.

Table 4. Precision, recall and F-Score of a deep neural network in the identification of academic biographies the web pages of academic researchers

	Precision	*Recall*	*F-Score*
CONTACTINFO	0.93	0.74	0.82
INTRODUCTION	0.76	0.86	0.80
PUBLICATION	0.95	0.95	0.95
RESINTERESTS	0.68	0.49	0.57

Based on the classification scheme (Kenekayoro, 2018), research interests can be part of a researchers biography. This is corroborated in the confusion matrix of the test set, where the majority of classification errors are instances belonging to research interests misclassified as introduction. Thus filtering out the publication category will arguably be sufficient for the extraction of biographies from academics' web pages. The high precision and recall of the publication class suggests that the learning model can be used to accurately exclude this category.

Representation of entity relationships

After extracting academic biographies from web pages and identifying the academic named entities using the techniques described in (Kenekayoro, 2018), the semantic relationships

between these entities may then be used to create a knowledge database that could be queried using semantic web technologies like SPARQL (Prud'hommeaux & Seaborne, 2008); a query language for matching RDF graphs (Pérez, Arenas, & Gutierrez, 2006). Sequence to sequence modelling that has been traditionally used for machine translation is a deep learning approach that may be applied to this task, but as the size of the available academic named entity dataset is limited, this is not investigated in this study.
Fig 1 shows the graphical representation of a textual biography converted to RDF triples / OWL. Fig 1 describes the steps used to identify entities and their relationships. The semantic connections between entities are determined by the verbs and prepositions that appeared close to these entities in a sentence.

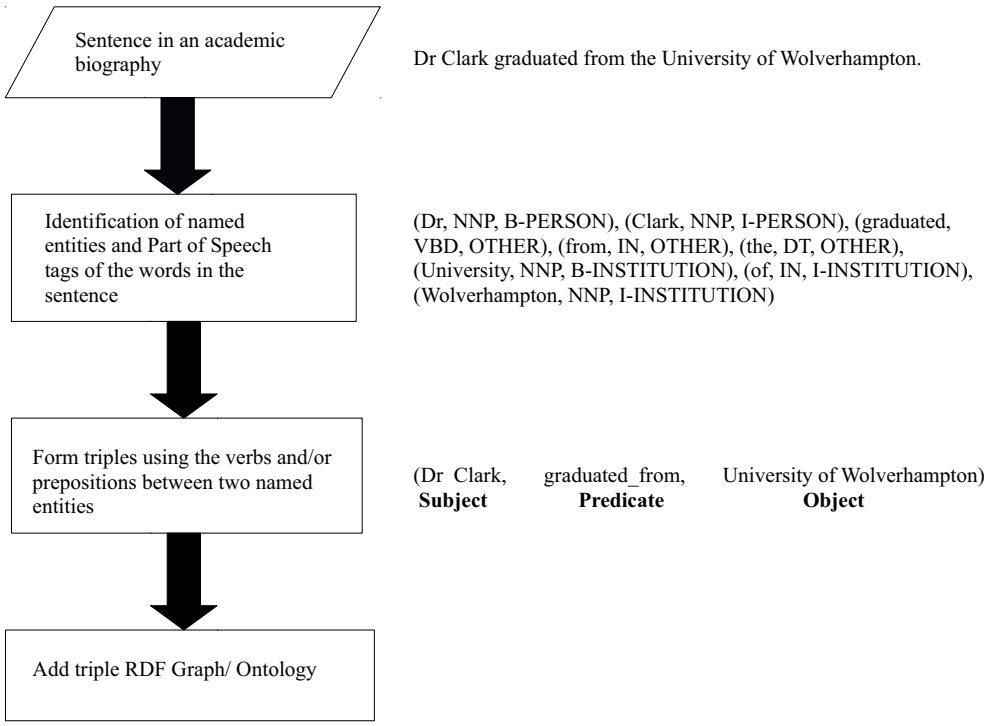


Fig 1. Flowchart describing the steps for extraction of semantic relations from a sentence in an academic biography.

Conclusions

Extracting semantic relations from web pages is important if web based data are to be used for Webometrics or Scientometrics studies. Part of the content of web pages are HTML tags and scripts whose purpose is solely for presentation, and some studies such as (Kenekayoro, 2018) may need only a particular section of a web page for its analyses. Thus, techniques that remove irrelevant content (HTML tags and scripts), as well as automatically annotate web page contents into their respective sections will be useful for webometric studies.
This study investigated the use of deep learning to extract academic biographies from the personal web pages of researchers. Deep learning has an advantage over traditional supervised learning algorithms in that domain knowledge for the identification of fine grained features is

not required, as relevant features are identified during the course of training a deep learning network. Nevertheless, deep learning may need considerably larger datasets compared to traditional machine learning methods in order to achieve generalisation.

The semantic relations between named entities in academic biographies can be used to create a knowledge base which can easily be processed and/or queried by computer algorithms, as opposed to plain text biographies. This study also demonstrated the use of verbs and prepositions that appeared in between named entities in a sentence to create a knowledge base of academic biographies in the form of RDF triples and/or OWL classes, making it possible to find new relations through logical inference and querying the knowledge base with semantic web technologies such as SPARQL.

Further work will aim to use the techniques described in this study to accurately reconstruct machine readable biographies from plain text biographies, which captures the activities and career of researchers in detail.

Acknowledgements

The author would like to thank the three anonymous reviewers for their useful comments.

References

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., et al. (2016). Tensorflow: a system for large-scale machine learning. *OSDI* (Vol. 16, pp. 265–283).

Bengio, Y., Simard, P., & Frasconi, P. (1994). Learning long-term dependencies with gradient descent is difficult. *IEEE transactions on neural networks*, *5*(3), 157–166.

Bird, S., Edward, L., & Ewan, K. (2009). *Natural language processing with Python: analyzing text with the natural language toolkit*. O'Reilly Media Inc.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32.

Caruana, R., & Niculescu-Mizil, A. (2006). An Empirical Comparison of Supervised Learning Algorithms. *Proceedings of the 23rd International Conference on Machine Learning*. Pittsburgh.

Cho, K., Van Merriënboer, B., Bahdanau, D., & Bengio, Y. (2014). On the properties of neural machine translation: Encoder-decoder approaches. *arXiv preprint arXiv:1409.1259*.

Cortes, C., & Vapnik, V. (1995). Support-Vector Networks. *Machine Learning*, *20*, 273; 273-297;

Culotta, A., Kanani, P., Hall, R., Wick, M., & McCallum, A. (2007). Author disambiguation using error-driven machine learning with a ranking loss function. *Sixth International Workshop on Information Integration on the Web (IIWeb-07), Vancouver, Canada*.

Cuxac, P., Lamirel, J.-C., & Bonvallot, V. (2013). Efficient supervised and semi-supervised approaches for affiliations disambiguation. *Scientometrics*, *97*(1), 47–58. Springer.

Fadlullah, Z., Tang, F., Mao, B., Kato, N., Akashi, O., Inoue, T., & Mizutani, K. (2017). State-of-the-art deep learning: Evolving machine intelligence toward tomorrow's intelligent network traffic control systems. *IEEE Communications Surveys & Tutorials*, *19*(4), 2432–2455. IEEE.

Freire, J., Koop, D., Santos, E., & Silva, C. T. (2008). Provenance for computational tasks: A survey. *Computing in Science & Engineering*, *10*(3). IEEE.

Gladun, A., Rogushina, J., Garcı´a-Sanchez, F., Martínez-Béjar, R., Fernández-Breis, J. T., & others. (2009). An application of intelligent techniques and semantic web technologies in e-learning environments. *Expert Systems with Applications*, *36*(2), 1922–1931. Elsevier.

Goh, A. T. C. (1995). Back-propagation neural networks for modeling complex systems. *Artificial Intelligence in Engineering*, *9*(3), 143–151. Elsevier.

Han, H., Giles, C. L., Manavoglu, E., Zha, H., Zhang, Z., & Fox, E. A. (2003). Automatic document metadata extraction using support vector machines. *Proceedings of the 3rd ACM/IEEE-CS joint conference on Digital libraries* (pp. 37–48).

Han, H., Giles, L., Zha, H., Li, C., & Tsioutsiouliklis, K. (2004). Two supervised learning approaches for name disambiguation in author citations. *Proceedings of the 2004 Joint ACM/IEEE Conference on Digital Libraries, 2004.* (pp. 296–305).

Hepp, M., Leymann, F., Domingue, J., Wahler, A., & Fensel, D. (2005). Semantic business process management: A vision towards using semantic web services for business process management. *e-Business Engineering, 2005. ICEBE 2005. IEEE International Conference* (pp. 535–540).

Hernández, I., Rivero, C. R., Ruiz, D., & Corchuelo, R. (2016). CALA: ClAssifying Links Automatically based on their URL. *Journal of Systems and Software*, *115*, 130–143. Elsevier.

Kenekayoro, P. (2018). Identifying named entities in academic biographies with supervised learning. *Scientometrics*. Retrieved from http://link.springer.com/10.1007/s11192-018-2797-4

Kenekayoro, P., Buckley, K., & Thelwall, M. (2014). Automatic classification of academic web page types. *Scientometrics*, *101*(2), 1–12. Springer Netherlands.

Kursa, M. B., & Rudnicki, W. R. (2010). Feature selection with the Boruta package.

Mamoshina, P., Vieira, A., Putin, E., & Zhavoronkov, A. (2016). Applications of deep learning in biomedicine. *Molecular pharmaceutics*, *13*(5), 1445–1454. ACS Publications.

Marcus, M., Kim, G., Marcinkiewicz, M. A., MacIntyre, R., Bies, A., Ferguson, M., Katz, K., et al. (1994). The Penn Treebank. *Proceedings of the workshop on Human Language Technology - HLT '94* (p. 114). Morristown, NJ, USA: Association for Computational Linguistics.

McGuinness, D. L., Van Harmelen, F., & others. (2004). OWL web ontology language overview. *W3C recommendation*, *10*(10), 2004.

Najafabadi, M. M., Villanustre, F., Khoshgoftaar, T. M., Seliya, N., Wald, R., & Muharemagic, E. (2015). Deep learning applications and challenges in big data analytics. *Journal of Big Data*, *2*(1), 1. Springer International Publishing.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., et al. (2011). Scikit-learn: Machine Learning in {P}ython. *Journal of Machine Learning Research*, *12*, 2825–2830.

Pennington, J., Socher, R., & Manning, C. D. (2014). GloVe: Global Vectors for Word Representation. *Empirical Methods in Natural Language Processing (EMNLP)* (pp. 1532–1543). Retrieved from http://www.aclweb.org/anthology/D14-1162

Pérez, J., Arenas, M., & Gutierrez, C. (2006). Semantics and Complexity of SPARQL. *International semantic web conference* (pp. 30–43).

Polidoro, F., Giannini, R., Conte, R. Lo, Mosca, S., & Rossetti, F. (2015). Web scraping techniques to collect data on consumer electronics and airfares for Italian HICP compilation. *Statistical Journal of the IAOS*, *31*(2), 165–176. IOS Press.

Prud'hommeaux, E., & Seaborne, A. (2008). SPARQL Query Language for RDF. W3C Recommendation, January 2008.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature*, *323*, 533–536.

Schmidhuber, J., & Hochreiter, S. (1997). Long short-term memory. *Neural Comput*, *9*(8), 1735–1780.

Shadbolt, N., Berners-Lee, T., & Hall, W. (2006). The semantic web revisited. *IEEE intelligent systems*, *21*(3), 96–101.

Sutskever, I., Vinyals, O., & Le, Q. V. (2014). Sequence to sequence learning with neural networks. *Advances in neural information processing systems* (pp. 3104–3112).

Tang, J., Alelyani, S., & Liu, H. (2014). Feature selection for classification: A review. *Data Classification: Algorithms and Applications*, 37.

Tang, J., Zhang, D., & Yao, L. (2007). Social network extraction of academic researchers. *Data Mining, 2007. ICDM 2007.* (pp. 292–301). IEEE.

Tang, J., Zhang, J., Yao, L., Li, J., Zhang, L., & Su, Z. (2008). Arnetminer: extraction and mining of academic social networks. *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 990–998). ACM.

Werbos, P. J. (1988). Generalization of backpropagation with application to a recurrent gas market model. *Neural networks*, *1*(4), 339–356. Elsevier.

Yosinski, J., Clune, J., Nguyen, A., Fuchs, T., & Lipson, H. (2015). Understanding neural networks through deep visualization. *arXiv preprint arXiv:1506.06579*.

Yu, K., Guan, G., & Zhou, M. (2005). Resume information extraction with cascaded hybrid model. *Proceedings of the 43rd Annual Meeting ofthe Association for Computational Linguistics* (pp. 499–506).

Zhang, G. P. (2000). Neural networks for classification: a survey. *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, *30*(4), 451–462.

University characteristics and probabilities for funding of proposals in the European Framework Programs

Fredrik Niclas Pirol[1], Lisa Scordato[2] and Dag W. Aksnes[3]

[1] fredrik.piro@nifu.no, [2] lisa.scordato@nifu.no, [3] dag.w.aksnes@nifu.no

NIFU, Nordic Institute for Studies in Innovation, Research and Education
Økernveien 9, 0653 Oslo, Norway

Abstract

This study is inspired by and a follow-up to the paper "Participations to European Framework Programs of higher education institutions and their association with organizational characteristics" by Lepori and colleages (Scientometrics, 2015) who studied participation of European higher education institutions (HEI) in projects funded under the European Framework Programs for Research and Innovation. By adding proposals to the analysis, we have investigated whether the findings of Lepori et al. – foremost that large and highly reputed universities are most involved in EU projects – is a matter of size and research intensity/reputation, or whether they are more successful in their proposal activities. Our findings indicate that size is not correlated with success rates, i.e. large HEIs participate in more projects due to their size, not because they are more successful. The academic reputation and research intensity of the HEIs do matter for success rates, but not after introducing previous success rates. Our interpretation of this is that many HEIs – irrespective of size – have acquired strong competences in partner selection, prioritization of calls to respond to, proposal writing, and a solid history of participation in the framework programmes, which makes them well equipped to get their proposals accepted.

Introduction

Large, highly-reputed universities have been shown in previous studies to dominate participations in the European Union's Framework Programmes for Research and Innovation (EU FPs), and EU FP participations is in general driven by a stable core of key actors, functioning as nodes in European R&D networks. In this study, we aim to provide new insight into characterizing European research as 'elite- driven'. The aim of our study is to investigate whether previous findings about determinants for participation are the same as the determinants for success in the proposal process in the EU FPs. The novelty of our study is the use of submitted proposals to the EU FPs, to study which characteristics of Higher Education Institutions (HEIs) that are associated with acceptance – and thereby funding – of the proposals, in a quantitative manner. The results of this study will thus add understanding to whether large and highly-reputed HEIs are involved in more projects because their large size enable them to submit many proposals, or whether their large overall degree of project participations is a result of these HEIs being relatively more successful in their proposal activities.

Former studies on participation in the EU FPs oriented at networks and participations

Although the literature on participations and involvement in EU FPs is comprehensive, little research has been performed concerning one of the main key words of EU FP participation: success rates. The success rate gives the ratio between the number of submitted proposals and the number of proposals that have been granted funding from the EU. The monitoring of the national success rates is key in many European countries' year-to-year follow-up on how their institutions are performing in the EU FPs. Existing studies have, to the best of our knowledge, relied upon project participation data in EU FPs only, and not on application data, thereby not taking the ratio between number of submitted proposals and number of granted projects (i.e. success rates) into account. With the possibility of using application data (including both rejected and successful proposals), comes the possibility to predict which institutional factors of the applicants that are associated with positive outcomes in the proposal review process.

Methodologically, the main problem of using proposals rather than projects, is the fact that no standardized database of proposal participants exist, while the project participant data is standardized by the European Commission, so that one can easily count the number of funded projects per institution. This is not so for the proposal data, and the work load of standardizing these data is an obvious impetus towards carrying out studies where success rates are used.

The strong emphasis in most EU calls on collaboration across countries and sectors (very often a requirement for funding), has sparked a vast literature on network studies within the EU FPs. The main findings from this literature is that EU FP participations are concentrated around a few key actors, and that the central nodes of the networks remain stable over time, indicating that past participations spur new participations, and that inclusion of new institutions into the EU FP projects do not change what has become a very stable set of participating partners. This pattern is found in HEIs as well as in publicly funded research organisations and in private companies (Breschi & Cusmano, 2004; Heringa, Hessels & van der Zouwen, 2016; Nokkola, Heller-Schuh & Paier, 2011; Okubo & Zitt, 2004; Paier & Scherngell, 2011; Roediger-Schluga & Barber, 2008). According to Hoenig (2017, p.4), members of the scientific community themselves "speculate that procedures of grant peer review might be subject to a massive 'Matthew effect'. Dynamics of cumulative advantage and disadvantage in gaining scientific recognition may partially explain a strong concentration of grants in only a few countries and institutions in Europe. Several mechanisms of such consolidation have been suggested. The most prominent is related to the scientific excellence or reputation of an institution. On a general note (not specifically related to EU FP funding), Frenken, Heimeriks & Hoekman (2017), using bibliometric data from the Leiden ranking, find that research performance differences among universities mainly stem from size, disciplinary orientation and country location. Large universities systematically over-perform in citation performance, international co-publications and in university-industry co-publications.

In studies of factors influencing peer review and funding decisions, such bibliometric indicators have been shown to be strongly associated with positive outcomes (Cabezas-Clavijo et al., 2013; Neufeld, Huber & Wegner, 2013; Vieira & Gomes, 2016), and lack of such associations may be explained by a self- (or pre-) selection of the 'inferior' organizations themselves (Neufeld & von Ins, 2011; Neufeld & Hornbostel, 2012), which has been found in the EU FPs on an overall level (Enger & Castellaci, 2016), and more specifically within the ERC (Neufeld, Huber & Wegner, 2013). Enger & Castellaci's (2016) study of Norwegian HEIs further showed that the propensity to apply is enhanced by prior participation in the EU FPs and the existence of complementary national funding schemes; furthermore, that the probability of succeeding is strengthened by prior participation as well as the scientific reputation of the applicant organization. The use – and importance – of bibliometric indicators such as field normalized citation indexes and highly cited papers – are also used for funding decisions in national research funding (e.g. Gunashekar, Wooding & Guthrie, 2017), with claims that the acquisition of research funds also at the individual level is more strongly correlated with the reputation of the research applicant than with the judgement of the proposal quality in the selection process (Hamann, 2016; Viner et al. 2006; Laudel 2006). Neufeld, Huber & Wegner (2013, p.245) argue that if nearly all actual applicants pass a certain threshold, other reasonable funding criteria may dominate the funding decision, most notably the quality and/or originality of the proposal (Melin & Danell, 2006), or whether the proposal is well-articulated (with a high quality of discourse) and/or with a topical overlap between the proposal references and the applicant's prior publications (Boyack, Smith & Klavans, 2018). Several studies have pointed at academic reputation and scientific productivity as strong markers of participating institutions in EU FPs

(Geuna, 1998; Henriques et al., 2009; Lepori et al. 2015; Nokkola, Heller-Schuh & Paier, 2011). In a recent study by Lepori et al. (2015), it was demonstrated that in a sample of 2,235 European HEIs, the level of participation in EU FP funded projects was strongly associated with organizational characteristics, particularly with size and academic reputation.

None of the EU-related studies above took the proposal activities of the institutions into consideration, they rather focused on the project participations. In a former study of success rates in the EU FPs, we have shown that bibliometric indicators were highly correlated with success rates, with the share of highly cited papers being the indicator most significantly correlated with success rates (Piro, Scordato & Aksnes, 2017). With respect to R&D collaboration, Lepori et al. (2015, p. 2153) argue that: "higher-reputed researchers and organizations will be sought to a greater extent as research partners and, therefore, move to the center of the network", but such an assumption – at the university level – does not take into account that the highly reputed institutions may be at the centre, i.e. participating in many collaborative projects, also because they simply submit many proposals, reflecting their large institutional size. Piro, Scordato and Aksnes (2016) investigated all EU FP consortia in EUs Seventh Framework Programme and Horizon 2020, involving participation from six selected European countries and concluded that a strong determinant for success in the EU FPs was previous participation in FP projects and a large network of European partners from past and ongoing projects. These findings are in concurrence with the 'centrality hypothesis', indicating that institutions that are already central in the European research network will become even more central as they are attractive partners to engage with for other institutions (Evans et al., 2011), thus strengthening the 'Matthew effect' of FP participation.

The 'Lepori study'

In the study *Participations to European Framework Programs of higher education institutions and their association with organizational characteristics*, Lepori et al. (2015) (from now on referred to as the Lepori study) analysed the patterns of participation of European HEIs to EU FPs, and their association with HEI characteristics. The sample included 2,235 HEIs in 30 countries in Europe, of which 861 had at least one EU FP participation in 2011.The concentration of participations was strong: a group of about 150 universities accounted for over 70 per cent of total participations in EU projects in 2011. For these HEIs institutional data from the European Tertiary Education Register (ETER) was matched with participations in EU FPs, using the EUPRO database. The methods of this study will be explained in detail in the methods section, as our analysis by large follow the design set out by Lepori and colleagues. The main results of the Lepori study were that EU FP participation is highly concentrated in a small group of HEIs that are characterized by a large volume of scientific publications, which combined with high citation rates, provide these HEIs with a strong (academic) reputation. The second main finding was that the participation of non-doctorate awarding HEIs in EU FPs is very limited, although they constitute for a very large share of the total student population. This finding is in line with other studies, emphasising the self-selection of institutions to EU FP participation (e.g. Enger & Castellaci, 2016). The number of participations tended to increase proportionally to organizational size and was strongly influenced by international reputation. We find this result particularly interesting, as it implies that there are no differences in the HEIs' ability to get their proposals submitted beyond size, i.e. the success rates are constant across the spectre of small and large HEIs. Finally, the authors found limited evidence of country effects on participation, or on the geographical distance from Brussels.

Methods

Our study draws upon three main data sources: 1) data about EU FP proposals and applications from the European Commission's data warehouse ECORDA, covering the years 2007-2017, 2) data from the Leiden ranking produced by Centre for Science and Technology Studies (CWTS) at Leiden University, and 3) data from the European Tertiary Education Register (ETER). We have used the November 2017 edition of ECORDA, which means that our FP7 data are complete, whereas the analysis of H2020 is restricted to the early results of that framework programme. In ECORDA, the institutional affiliations of applicants are not standardized. It is therefore not possible to use the data for calculating success rates without going through the process of standardizing the institutional names. The novelty of our study is the build-up of a completely new database for participation in both EU proposals and projects. We have standardized all institutional addresses in EU's application and project databases of FP7 and H2020 (approximately 1.4 million institution names). To the best of our knowledge, such a standardized data file including both applicants and grant recipients does not exist anywhere else. This has enabled us to extract data about all institutions' total volume of applications and projects, thereby making it possible to calculate success rates for all institutions in the database.

The European HEIs identified in ECORDA were then matched with ETER-data and citation data from the Leiden ranking, so that we could calculate the same HEI indicators as used in the Lepori study. ETER currently includes 2,764 HEIs in the 36 considered countries, but data on the relevant indicators is not available for all countries, or for all HEIs in each country. The Leiden ranking includes data on the publication output of the world's largest universities. All universities in the Leiden ranking have been given a value representing their mean number of fractionalized scientific publications during the years 2011-2014 (the Leiden ranking does not provide numbers further back in time). The citation impact of the universities is based on their average field normalized citation indicator during the same period (see Waltman et al., 2012).

Independent variables

The following HEI indicators were used in the Lepori study as independent variables, and our calculations and data sources of these variables follows the same methodology, with one exception – the Reputation indicator. Lepori and colleagues measured reputation as the product between the normalized impact factor and the total number of publications from the university (cf. van Raan, 2008), normalized by the number of academic staff (which is an indicator that is missing in many countries in ETER). Our reputation indicator differs in three ways: first, we have used bibliometric data from the Leiden ranking rather than the SCIMAGO Institutional ranking. Second, we have not divided the scores by the number of academic staff as this indicator is already included in the analysis as a separate indicator.

Third, our *reputation indicator* is made up by an average over four years, rather than from one unique year. In the Leiden ranking, many universities are excluded since they do not have a large enough publication volume to meet the inclusion threshold. Here, they will be treated as a separate dummy variable, rather than representing missing values. This is reasonably justified by studies documenting the correlation between publication size and publication impact (e.g. Frenken, Heimeriks & Hoekman, 2017). The remaining independent variables, follow from the same data sources and calculating methods as in the Lepori study: *Size* – the number of academic staff is measured by full time equivalents (FTEs). *Research intensity* is the ratio between the share of PhD graduates and the number of undergraduate graduates. *Teaching load* is a measure of the HEI's orientation towards education. It is computed as the ratio between the total number of undergraduate students and academic staff. *External funding* is the share of third-party funds of total HEI revenues.

Subject specialization was included in the Lepori study because EU FP funding is concentrated on technological domains, with fewer funding options for scientists within social sciences and humanities (SSH). European HEIs display different subject composition and many, mostly either in technology or towards humanities, are highly specialized (Lepori, Probst & Baschung, 2010), making the subject specialization of the HEIs a plausible factor in explaining the overall volume of EU FP participations. Lepori and colleagues therefore expected that a stronger orientation towards fields like ICT and engineering would be associated with a higher number of participations, whereas an orientation towards SSH would have a lower number. Such an assumption is much more intuitive when studying the number of participations compared to success rates, because even HEIs with a low number of submitted proposals may in fact be very successful (and vice versa). Several studies have found differences by subject specialization/profiles in both productivity and citation impact, thus adding support to such a factor as a determinant of institutional success (Dundar & Lewis, 1995; Lopez-Ilescas, Moya-Anegon & Moed, 2011; Moed, Moya-Anegon, Lopez-Ilescas & Visser, 2011; Piro, Aksnes & Rørstad, 2013). Two variables for subject specialization was used in the Lepori study. First, a set of dummies for specialized HEIs, identifying the HEIs that had more than half of their undergraduate students in a field. Lepori and colleagues used the standard classification of Fields of Education adopted in international educational statistics. Second, the share of undergraduate students in natural and technical sciences (including ICT).

The importance of past EU FP experience: Further, we explore the importance of past proposals, projects – and success rates. In a study of research consortia in six European countries (Piro, Scordato & Aksnes, 2016), it was found that the institutions with the highest success rates formed consortia with other partners that fitted their own profiles well, e.g. substantial experience from past EU FP participations. Hence, universities that are on the inside of the EU FP networks benefit from their institutional reputation, their know-how on how to write proposals and their networks that they can draw upon. The importance of strengthening one's degree of centrality in the European research network, has even been highlighted by the European Commission (2015, p.112) as an importance driver towards success, simply recommending that institutions should make efforts in increasing their networks: "Stamina, repeated participation, and a willingness to increase one's connections are the only way forward to better one's position when on the periphery". When studying the probability for having a proposal accepted in H2020, we believe studying the number of proposals, projects and success rates (from the equivalent programs) under FP7 are important variables. Proposals indicate the experience (and learned knowledge) from participating in writing of proposals, projects indicate the experience (and learned knowledge) from successfully completing EU-funded projects, and success rates indicate the ability pick the right partners, submit proposals to relevant calls, be in possession of the necessary scientific quality etc.

Two of the HEI indicators from the Lepori study have been left out in our study for pragmatic reasons. The indicator Geographical Distance which measured as the distance between the HEI's headquarters (as identified by their geographical coordinates in ETER) and Brussels, and was not statistically significant in any models, and is not included here. Another HEI variable has been ex-post excluded as we found no correlation at all with success rates, namely whether, or not, the HEI is a PhD awarding institution. While this factor is strongly associated with the propensity to apply, and with the number of projects (i.e. low numbers for non-awarding institutions), it had no effect in our study, as indicated by the fact that only 9,768 (2.4 per cent) out of 412,636 proposal participations in the period 2007-2017 were from HEIs that do not award PhDs. The Lepori study also used two country indicators: Higher Education Research and Development Expenditures (HERD) per inhabitant in purchasing power parities, and

dummy variables for 1) new member states, and 2) associated countries. The former includes as unevenly comparable countries as Switzerland and the Former Yugoslav Republic of Macedonia. We have rather chosen to have dummy variables for each country.

Study sample

In ECORDA we have identified 4,957 HEIs from a total of 174 countries that in the period 2007-2017 were involved in one or more proposals submitted to the EU FPs. Of these, 1,880 were European (Table 1). Luxembourg and Liechtenstein are represented with one HEI each, and is therefore not included in Table 1, as legal restrictions permit us from showing application data that can be linked to one specific institution. Germany and France have by far the highest amount of HEIs in ECORDA (278 and 213 respectively). The largest volume of the proposal participations is from the UK, whose HEIs account for 20.6 per cent of the proposal participations. This is much higher than second-placed Germany (11.3 per cent) and the other largest nations: Italy (10.2 per cent), Spain (7.2 per cent), Netherlands (5.8 per cent), Sweden (4.8 per cent) and France (4.6 per cent). The French case is very special and is arguably caused by the methodological challenges appearing in France when the proposals have been submitted in the name of the National Center for Scientific Research (CNRS) or one of the many laboratories funded by the CNRS or by other national funding agencies in France such as the National Institute for Health and Medical Research (INSERM). For this reason, and data unavailability for French HEIs in ETER, the French HEIs are not part of our analysis.

Table 1. Distribution of Higher Education Institutions across countries in EU FPs 2007-2017.

		Submitted proposals				Projects				
Country	HEIs (N)	N	Mean	Min	Max	N	Mean	Min	Max	Success rate
Austria	57	11407	200.1	1	1959	1801	31.6	0	322	15.8
Belgium	30	12903	430.1	1	4798	2120	70.7	0	812	16.4
Bulgaria	38	1537	40.4	1	356	214	5.6	0	65	13.9
Croatia	20	1868	93.4	1	1267	211	10.6	0	154	11.3
Cyprus	14	2628	187.7	1	1311	358	25.6	0	172	13.6
Czech Republic	37	5013	135.4	1	1318	709	19.2	0	180	14.1
Denmark	23	12270	533.5	1	3670	2184	95.0	0	697	17.8
Estonia	9	2002	222.4	3	1000	295	32.8	0	165	14.7
Finland	40	10680	267.0	3	2482	1418	35.5	0	384	13.3
France	213	19164	90.0	1	1354	3354	15.7	0	268	17.5
Germany	278	46709	168.0	1	2268	8160	29.4	0	483	17.5
Greece	39	12766	327.3	1	1921	1687	43.3	0	298	13.2
Hungary	37	4263	115.2	1	914	662	17.9	0	141	15.5
Iceland	6	690	115.0	9	514	87	14.5	0	74	12.6
Ireland	25	10472	418.9	4	1943	1735	69.4	0	352	16.6
Italy	109	42036	385.7	1	2796	5431	49.8	0	443	12.9
Latvia	25	1199	48.0	1	424	202	8.1	0	78	16.8
Lithuania	30	1865	62.2	1	522	252	8.4	0	75	13.5
Macedonia	7	416	59.4	2	333	56	8.0	0	45	13.5
Malta	2	535	267.5	22	513	84	42.0	4	80	15.7
Montenegro	4	151	37.8	1	120	26	6.5	0	26	17.2
Netherlands	42	23749	565.5	1	3369	4454	106.0	0	640	18.8
Norway	30	5612	187.1	1	1693	846	28.2	0	243	15.1

Poland	148	8422	56.9	1	851	1160	7.8	0	149	13.8
Portugal	53	7745	146.1	1	1733	1064	20.1	0	306	13.7
Romania	70	4184	59.8	1	660	382	5.5	0	66	9.1
Serbia	15	1600	106.7	1	782	180	12.0	0	92	11.3
Slovakia	28	1764	63.0	1	360	218	7.8	0	46	12.4
Slovenia	25	2957	118.3	1	1872	324	13.0	0	231	11.0
Spain	75	29797	397.3	1	2164	3979	53.1	0	374	13.4
Sweden	32	19587	612.1	3	2671	3368	105.3	0	478	17.2
Switzerland	25	15107	604.3	1	3290	3066	122.6	0	779	20.3
Turkey	144	6014	41.8	1	656	760	5.3	0	103	12.6
UK	148	84961	574.1	1	4753	15628	105.6	0	1099	18.4
Total	**1880**	**412636**	**218.6**	**1**	**4753**	**66558**	**35.3**	**0**	**1099**	**16.1**

The highest success rates are found in Switzerland (20.3 per cent), the Netherlands (18.8) and the UK (18.4 per cent). Some of the smaller nations stand out with high success rates, such as Denmark (17.8 per cent) with a relatively large volume of submitted proposals, and Montenegro (17.2 per cent) with a very low volume of submitted proposals. Among the 36 countries listed in Table 1, ten have been left out of our study due to data limitations in ETER (Austria, Estonia, France, Greece, Iceland, Macedonia, Montenegro, Romania, Slovenia and Turkey). Seven countries do not have any HEIs that are listed in the Leiden ranking, and all institutions from these countries are thus represented by a dummy indicator.

Table 2. Descriptive statistics on independent variables for HEIs (n=1,307).

Indicator	Valid (N)	Missing (N)	Mean	St.dev	Skew-ness	Min.	1Q	Median	3Q	Max
PUB	232	1075	3001.8	1950.5	1.660	955.6	1471.2	2360.6	4014.8	11830.5
CIT	232	1075	1.0196	0.1849	0.028	0.540	0.915	1.011	1.145	1.570
Reputation	232	1075	3247.9	2634.5	2.298	603.8	1395.6	2302.0	4268.8	17673.5
Ac.Staff FTE	1246	60	763.8	1013.1	2.532	6.8	136.4	364.0	985.0	7083.8
Res. intensity	784	523	0.0803	0.5272	24.984	0.00	0.0126	0.0308	0.0630	14.89
Teaching load	1176	133	0.1149	0.3472	18.237	0.01	0.0484	0.1149	0.1618	8.49
3rd Party %	802	505	0.1274	0.1219	2.082	0.00	0.0442	0.0886	0.1743	0.88
NATTECH	1260	47	25.93	26.89	0.063	0.00	2.80	19.23	36.28	100.00
FP7 proposals	1307	0	141.90	311.63	3.959	1.00	4.00	18.00	116.75	2744.00
FP7 projects*	915	0	25.81	66.14	5.108	0.00	0.00	2.00	18.75	734.00
FP7 success rate*	915	0	0.1339	0.156	2.793	0.00	0.00	0.125	0.190	1.00

*Only for HEIs that did get proposals accepted

Table 2 shows the HEI values on the independent variables of our study. Among the 1,307 HEIs from the 25 selected countries, only 232 have sufficient publication numbers (thereby also citation indexes) to appear in the Leiden ranking. The most skewed indicator is research intensity, where several universities have the value 0. Most variables are highly skewed, and in the full sample (FP7 and H2020 combined), the HEIs have contributed with 352,322 proposal contributions, of which 57,703 (16.4 per cent) were successfully funded. This distribution of responses to rejected/accepted is suitable for logistic regression. The dummy variables for subject specialization (Table 3) identifies groups of specialized HEIs in individual fields, which includes 40 per cent of our sample. Most specialized HEIs are in Business, administration and

law (139), in ICT, engineering, manufacturing and construction (105) and in Arts and humanities (102).

Table 3. Distribution of specialized HEIs.

Fields of specialization	N	% of included HEIs
Education	27	2.1
Arts and Humanities	102	7.8
Social sciences, journalism and information	32	2.5
Business, administration and law	139	10.6
Natural sciences, mathematics and statistics	2	0.2
ICT, engineering, manufacturing and construction	105	8.0
Agriculture, forestry, fisheries and veterinary	13	1.0
Health and welfare	60	4.6
Services	37	2.5
Specialized universities in total	517	39.6

Results

We use logistic regression analysis, displaying the results by odds ratios (with 95% confidence intervals), i.e. the probability of having a proposal accepted. The analysis is conducted in two steps. We first include the same indicators that were used by Lepori et al. (2015), and then limit the analysis to H2020 only, with HEIs' number of submitted proposals, funded projects and success rates from FP7 as independent variables. Given the substantially lower numbers of proposals submitted to H2020 in our dataset, fewer significant results are expected to be found here. No specialization is the reference category for subject specialization. In the reputation indicator, universities that are not included in the Leiden ranking is the reference category. For the country dummies, the UK is the reference category (and we only show statistically significant country differences in the tables). For all other indicators, the reference category is the indicator with the lowest value.

In the first model of Table 4, we find that HEIs specialized in the technical fields (ICT, engineering, manufacturing and construction) and in the social sciences (including journalism and information sciences) have a statistically significant higher probability for having their proposals accepted for funding. The first finding may not be unexpected, given the technological profile of the EU FPs, while the finding for social sciences at first may be so. However, it is important to note that this does not contradict the opposite finding in the Lepori study. HEIs specialized in the social sciences do have a lesser propensity to be involved in EU FP projects compared to other HEIs, but at the same time, in the thematic areas where they do submit proposals, they are seemingly successful in getting funding, although at a much lower level/volume compared to e.g. technical universities. Also, with the subject specialization indicator already in the model, we find no significant differences for the indicator percentage of students within natural sciences and technology.

The reputation indicator does not show a linear relationship, but the top three groups of HEIs with seemingly high reputation scores, have odds significantly higher than in the reference group (which represent most European universities, with a publication volume below the Leiden ranking's threshold level) and in the group of the smallest and least reputed HEIs in the Leiden ranking. The confidence interval of the most reputed group (4) indicates that there is a higher probability of funding compared to groups 3-4, although not statistically significant. The

most research intensive HEIs are statistically different from the least research intensive HEIs, and there is a statistically significant association between low teaching load and increased probability for funding. Size of the HEI (by FTEs) and the third-party funding indicator are not statistically significant. Several of the country dummies, show statistical low probability for funding compared to the UK.

We then add experience from FP7 to the model and limit the outcomes to H2020. With a smaller sample, and with many more variables entered to the model, it is no surprise that many of the previously significant estimates now become non-significant. What remains statistically significant must be claimed to be robust findings. The following changes in the model should be emphasized. First, the level of submitted proposals in FP7 does not seem to have any influence on the probability for having proposals in H2020 accepted. Second, having experience from involvement in many, compared to fewer, projects funded in FP7 is significantly associated with higher odds for acceptance in H2020. Third, although not perfectly mutually exclusive (based on confidence intervals) the HEIs' success rates in FP7 is statistically significant to positive funding decisions in H2020 (i.e. a gradual increase from the least to the most successful HEIs in FP7). Our interpretation of this is that success is not based on scaling effects from a large volume of submitted proposals, but rather on some HEIs being very good at pinpointing the right calls and creating strong proposals, thus leading to high success rates.

Table 4. Logistic regression analysis (odds ratios and 95% CI) for associations between university characteristics and proposal acceptance in FP7/H2020.

	FP7 and H2020		H2020 only	
	OR	95% CI	OR	95% CI
Agriculture, Forestry, Fishery, Vet.	1.409	0.949 – 2.092	*2.184	1.162 – 4.104
Arts, Humanities	0.969	0.642 – 1.463	1.137	0.601 – 2.151
Business, Law	1.073	0.937 – 1.229	0.958	0.760 – 1.208
Education	0.648	0.340 – 1.235	0.532	0.124 – 2.277
ICT, Engineering, Manuf., Constr.	***1.113	1.068 – 1.160	1.063	0.987 – 1.145
Health, Welfare	0.972	0.880 – 1.074	0.910	0.760 – 1.090
Natural Sciences, Math, Statistics	*1.765	1.029 – 3.025	1.369	0.520 – 3.603
Services	1.280	0.642 – 2.552	0.865	0.199 – 3.758
Social Sciences, Journalism, Inform.	***1.460	1.255 – 1.697	***1.785	1.388 – 2.295
Reputation (1)	0.953	0.892 – 1.017	0.952	0.849 – 1.068
Reputation (2)	***1.131	1.069 – 1.197	1.052	0.946 – 1.171
Reputation (3)	***1.134	1.072 – 1.200	1.055	0.943 – 1.181
Reputation (4)	***1.244	1.176 – 1.316	1.005	0.892 – 1.134
Academic Staff FTEs (1)	1.130	0.807 – 1.580	1.288	0.672 – 2.470
Academic Staff FTEs (2)	1.250	0.898 – 1.741	1.604	0.842 – 3.055
Academic Staff FTEs (3)	1.243	0.892 – 1.732	1.632	0.856 – 3.112
Research Intensity (1)	1.078	0.996 – 1.165	0.943	0.824 – 1.079
Research Intensity (2)	***1.176	1.091 – 1.268	0.986	0.866 – 1.123
Research Intensity (3)	***1.198	1.104 – 1.299	0.940	0.816 – 1.083
Teaching Load (high)	0.994	0.931 – 1.061	1.000	0.898 – 1.114
Teaching Load (medium)	1.004	0.933 – 1.081	1.004	0.887 – 1.137
Teaching Load (low)	*1.079	1.002 – 1.163	1.078	0.950 – 1.223
Third Party Funding (1)	0.944	0.883 – 1.010	***0.849	0.757 – 0.953
Third Party Funding (2)	0.979	0.913 – 1.050	0.916	0.813 – 1.032
Third Party Funding (3)	1.023	0.951 – 1.101	0.900	0.794 – 1.019
Nat-Tech per cent (1)	1.039	0.943 – 1.146	1.091	0.925 – 1.286
Nat-Tech per cent (2)	1.013	0.918 – 1.119	1.095	0.927 – 1.294
Nat-Tech per cent (3)	1.060	0.958 – 1.172	1.019	0.858 – 1.210
FP7 Number of proposals (1)			0.992	0.882 – 1.115
FP7 Number of proposals (2)			0.988	0.852 – 1.145
FP7 Number of proposals (3)			1.054	0.882 – 1.260

FP7 Number of Projects (1)			1.076	0.952 – 1.216
FP7 Number of Projects (2)			*1.172	1.007 – 1.365
FP7 Number of Projects (3)			*1.209	1.007 – 1.453
FP7 Success rate (1)			***1.192	1.102 – 1.289
FP7 Success rate (2)			***1.242	1.137 – 1.355
FP7 Success rate (3)			***1.336	1.214 – 1.471
Constant	0.162		***0.053	
Belgium	***0.845	0.792 – 0.901	**0.844	0.754 – 0.946
Cyprus	0.936	0.787 – 1.113	**1.581	1.184 – 2.112
Denmark	***0.905	0.858 – 0.954	0.986	0.902 – 1.077
Germany	***0.932	0.897 – 0.969	1.053	0.978 – 1.134
Italy	***0.732	0.696 – 0.770	0.911	0.827 – 1.003
Lithuania	**0.827	0.715 – 0.957	0.875	0.681 – 1.123
Netherlands	***0.938	0.897 – 0.980	0.967	0.896 – 1.045
Norway	***0.828	0.765 – 0.895	0.991	0.867 – 1.132
Portugal	***0.774	0.717 – 0.836	1.135	0.996 – 1.293
Slovakia	***0.680	0.576 – 0.803	*0.722	0.543 – 0.960
Sweden	***0.875	0.836 – 0.916	0.950	0.872 – 1.034
-2 Log likelihood	253387		94171	
Nagelkerke R Square	0.0102		0.008722	
Chi-square	***1719		***568	
Rejected proposals (N)	231.466		101.231	
Accepted proposals (N)	47.645		16.328	
Total proposals (N)	279.111		117.559	

*p<0.05, **p<0.01, ***p<0.001

Because of the introduction the FP7 indicators, there are now very few other covariates that are statistically significant. Most notably, the reputation and research intensity indicators are no longer significant. The HEIs specialized in the natural sciences no longer have higher odds for funding, but the association found for specialization in social sciences is even stronger in this model. It is also interesting that the HEIs specialized in agriculture, forestry, fishery and veterinary sciences now have a statistically higher chance for funding compared to the non-specialized HEIs.

Discussion

The starting point of our study was Lepori et al.'s (2015) study which has served as inspiration and methodological framework to us. Since the two studies are investigating very different research questions (number of participations versus probability for funding acceptance), any contradicting conclusions do not imply any sort of methodological criticism of the Lepori study on our behalf. It is nevertheless interesting to compare some of the seemingly competing findings between the two studies. Both studies support the hypothesis that past results are decisive for future results. The two studies operate, however, with different explanations to this. The Lepori study highlighted the network structure of the EU FP participants, where "current participations to EU-FP largely generate new ones, because they are borne from existing collaborative links. At the same time, new participations are associated with HEI reputation and, therefore, highly-reputed HEIs move to the center of the EU-FP network until an equilibrium state is reached. In other words, reputational mechanisms maintain the stability of the network ensuring that consistently the higher reputed HEIs are found in its core and that, in the long term, an increase in reputation will also lead to larger number of EU-FP participations" (p.2174). Therefore, the large and reputed HEIs will be involved in many projects. Our contribution to this finding is that there were no scaling-effects of size on probability for funding acceptance, simply indicating that large and reputed HEIs are involved in more projects because they submit more proposals. This does not mean, however, that the large and reputed

HEIs more often get their proposals funded. Size (by FTEs) was never associated with funding decisions in our study, while academic reputation was until we controlled for success rate in FP7. Our interpretation of this is that some of the most successful HEIs in the EU FPs do not submit as many proposals as their size would indicate, rather they effectively submit proposals to calls where they have good chances for funding acceptance. Like Lepori et al. (2015) we found no effect of research intensity and follow their argumentation that this can be explained by the fact that most of such an effect is absorbed by reputation.

Despite being based on a very large data set, our regression models suffer from low explanatory power. We believe the model fit could improve if more fine-tuned analysis had been carried out. The results so far (but will in the journal paper) do not distinguish between different pillars/sub-programmes under the FPs, where university characteristics may be of different importance (most notably ERC vs. programs for SMEs). Further, we did not distinguish between the coordinator and the partner role of the proposals (or projects). It seems plausible that the reputation or impact (or any other characteristics) of the coordinating university will be more important than those of the other participants involved in the consortium. In a previous study, using university ranking data, we found that the success rates for universities differ and depend on whether they have coordinated or contributed as partner in a proposal (Piro, Scordato & Aksnes, 2017), indicating that the 'quality' of the coordinator is given more weight in the review process of proposals than for the partners in some programmes.
Finally, we would like the stress the importance of the partners involved in the project proposals. While this not important in the ERC and MSCA, it is certainly so in the collaboration programs where the EU FP track record of the industrial partners may be more decisive for the funding decision that the academic characteristics of the HEI partners in the consortia.

References

Boyack, K.W., Smith, C. & Klavans, R. (2018). Toward predicting research proposal success. *Scientometrics,* 114, 449-461.

Breschi, S. & Cusmano, L. (2004). Unveiling the texture of a European Research Area: emergence of oligarchic networks under EU Framework Programmes. *International Journal of Technology Management*, 27(8), 747-772.

Cabezas-Clavijo, A. Robinson-Garcia, N., Escabias, M & Jiménez-Contreras, E. (2013). Reviewers' Ratings and Bibliometric Indicators: Hand in Hand When Assessing Over Research Proposals? *PLOS ONE,* 8(6), e68258.

Dundar H. & Lewis, D.R. (1995). Departmental productivity in American universities: economies of scale and scope. *Economics of Education Review,* 14(2), 119-144.

Enger, S.G. & Castellaci, F. (2016). Who gets Horizon 2020 research grants? Propensity to apply and probability to succeed in a two-step analysis. *Scientometrics,* 109, 1611–1638.

European Commission (2015). *Study on Network Analysis of the 7th Framework Programme Participation. Final report.* Brussels: European Commission, Directorate General for Research & Innovation.

Evans, T.S., Lambiotte, R. & Panzarasa, P. (2011). Community structure and patterns of scientific collaborations in business and management. *Scientometrics,* 89(1), 381-396.

Frenken, K., Heimeriks, G.J. & Hoekman, J. (2017). What drives university research performance? An analysis using the CWTS Leiden Ranking data. *Journal of Informetrics*, 11, 859-872.

Geuna, A. (1998): Determinants of university participation in EU-funded R&D cooperative projects. *Research Policy*, 26(6), 677-687.

Gunashekar, S., Wooding, S. & Guthrie, S. (2017). How do NIHR peer review panels use bibliometric information to support their decisions? *Scientometrics*, 112, 1813-1835.

Hamann, J. (2016). The visible hand of research performance assessment. *Higher Education*, 72, 761-779.

Henriques, L., Schoen, A. & Pontikakis, D. (2009). *Europe's top research universities in FP6: scope and drivers of participation.* European Commission: JRC Technical Notes (Vol. 53681).
Heringa, P.W., Hessels, L.K. & van der Zouwen, M. (2016). The Dynamics of the European water research network: a diversifying community with a stable core. *Water Policy,* 18, 493-512.
Hoenig, B. (2017). *Europe's New Scientific Elite. Social Mechanisms of Science in the European Research Area*. London: Routledge.
Laudel, G. (2006). The art of getting funded: How scientists adapt to their funding conditions. *Science and Public Policy*, 33(7), 489-504.
Lepori, B., Probst, C. & Baschung, L. (2010): Patterns of subject mix of higher education institutions: A first empirical analysis from the AQUAMETH database. *Minerva,* 48(1), 73-99.
Lepori, B., Veglio, V., Heller-Schuh, B., Scherngell, T. & Barber, M. (2015): Participations to European Framework Programs of higher education institutions and their association with organizational characteristics. *Scientometrics,* 105(3), 2149-2178.
Lopez-Ilescas C., de Moya-Anegon, F. & Moed, H. (2011). A ranking of universities should account for differences in their disciplinary specialization. *Scientometrics*, 88(2), 563-574.
Melin, G. & Danell, R. (2006). The Top Eight Percent: Development of Approved and Rejected Applicants for a Prestiguous Grant in Sweden. *Science and Public Policy*, 33(10), 702-712.
Moed H., de Moya-Anegon, F., Lopez-Ilescas, C. & Visser, M. (2011). Is concentration of university research associated with better research performance? *Journal of Informetrics*, 5(4), 649-658.
Neufeld, J. & Hornbostel, S. (2012). Do the 'Best Apply? *Research Evaluation*, 21(3), 1-10.
Neufeld J. & von Ins, M. (2011). Informed Peer Review and Uninformed Bibliometrics. *Research Evaluation*, 20(1), :31-46.
Neufeld, J., Huber, N. & Wegner, A. (2013). Peer review-based selection decisions in individual research funding, applicants' publication strategies and performance: The case of the ERC Starting Grants. *Research Evaluation*, 22, 237-247.
Nokkola, T., Heller-Schuh, B. & Paier, M. (2011). "Ranking lists and European Framework Programmes: Does university status matter for performance in Framework Programmes"? In P. T. D. Dill (Ed.): *Public Vices, Private Virtues? Assessing the Effects of Marketization in Higher education* (pp. 111-139). Rotterdam: Sense Publishers.
Okubo, Y. & Zitt, M. (2004). Searching for research integration across Europe: a closer look at international and inter-regional collaboration in France. *Science and Public Policy*, 31(3), 213-226.
Paier, M. & Scherngell, T. (2011). Determinants of Collaboration in European R & D Networks: Empirical Evidence from a Discrete Choice Model. *Industry and Innovation*, 18(1), 89-104.
Piro, F.N., Aksnes, D.W. & Rørstad, K. (2013). A Macro Analysis of Productivity Differences Across Fields: Challenges in the Measurement of Scientific Publishing. *Journal of the American Society for Information Science and Technology (JASIST),* 64(2),307-20.
Piro, F.N., Scordato, L. & Aksnes, D.W. (2016). *Choosing the right partners. Norwegian participation in European Framework Programmes*. NIFU report 41/2016. Oslo: NIFU.
Piro, F.N., Scordato, L. & Aksnes, D.W. (2017). The association between university research indicators and success rates in the European Framework Programmes. *Proceedings of the 16th International Conference on Scientometrics & Informetrics (ISSI*), Wuhan, China.
Roediger-Schluga, T. & Barber, M. (2008). R&D collaboration networks in the European Framework Programmes: data processing, network construction and selected results. *International Journal of Foresight and Innovation Policy,* 4(3-4), 321-347.
Van Raan, A.F.J. (2008). Bibliometric statistical properties of the 100 largest European universities: Prevalent scaling rules in the science system. *Journal of the American Society for Information Science and Technology*, 59(3), 461-475.
Vieira, E. & Gomes, J.A.N.F. (2016). The bibliometric indicators as predictors of the final decision of the peer review. *Research Evaluation,* 25(2), 170-183.
Viner, N., Green, R. & Powell, P. (2006). Segmenting academics: Resource targeting of research grants. *Science and Public Policy*, 33(3), 166-178.
Waltman, L., Calero -Medina, C., Kosten, J., Noyons, E., et al. (2012). The Leiden Ranking 2011/2012: Data collection, indicators, and interpretation. *Journal of the American Society for Information Science and Technology,* 63(12), 2419-2432.

The Integrated Impact Indicator (*I3*) and the Journal Impact Factor: A Non-Parametric Alternative

Loet Leydesdorff,*[,1] Lutz Bornmann,[2] and Jonathan Adams[3]

[1] loet@leydesdorff.net
Amsterdam School of Communication Research, University of Amsterdam, PO Box 15793, 1001 NG Amsterdam, The Netherlands;

[2] bornmann@gv.mpg.de
Max Planck Society, Administrative Headquarters, Hofgartenstr. 8, 80539 Munich, Germany;

[3] jonathan.adams@kcl.ac.uk
The Policy Institute at King's, King's College London, 22 Kingsway, London, WC2B 6LE, UK; Institute for Scientific Information, Clarivate Analytics, Blackfriars Road, London, UK.

Abstract

We elaborate the integrated impact indicator (*I3*) and apply it to more than 10,000 journals and compared the results with other journal metrics. *I3* provides a non-parametric statistic combining the measurement of impact in terms of citations (*c*) and performance in terms of publications (*p*) into a single metric. We argue for weighting using four percentile classes: the top-1% and top-10% as excellence indicators; the top-50% and bottom-50% as performance indicators. Unlike the *h*-index, which also appreciates both *c* and *p*, *I3* correlates significantly with *both* the total number of citations and publications. *I3* and *h*-index are size-dependent; division of *I3* by the number of publications (*I3/N*) provides a size-independent indicator which correlates strongly with (the standard two-year) impact factor (*JIF*2) and *JIF*5 (*JIF* based on five publication years). The values of *I3* and *I3/N* can be statistically tested against the expectation or one another using chi-square tests or effect sizes. Thus, one can indicate both performance and impact with a single number. We provide a template (in Excel) online for relevant tests.

Introduction

Citations create links between publications; but to relate citations to publications as two different things, one needs a model (for example, an equation). *JIF*2, for example, is defined as the number of citations in the current year (*t*) to any of a journal's publications of the two previous years (*t*-1 and *t*-2), divided by the number of citable items (substantive articles, reviews, and proceedings) in the same journal in these two previous years. *JIF*2 provides a functional approximation of the mean early citation rate per citable item. However, the mean of a skewed distribution provides less information than the median as a measure of central tendency. To address this problem, McAllister *et al*., (1983, at p. 207) proposed the use of percentiles or percentile classes as a non-parametric indicator (Narin, 1987; see later: Bornmann & Mutz, 2011; Tijssen, Visser, & van Leeuwen, 2002). On the basis of a list of criteria provided by Leydesdorff, Bornmann, Mutz, & Opthof (2011), two of us have developed an Integrated Impact Indicator (*I3*) based on the integration of the quantile values attributed to each element in a distribution (Leydesdorff & Bornmann, 2011). In this study, we further develop an *I3*-type journal indicator that combines both publications and citations in a single number.

An approach based on percentiles can be considered a "second generation indicator" for two reasons. First, one builds on the first-order approach that Garfield (1979, 2003, 2006) developed for the selection of journals. Second, the objective of journal selection is very different from the

purposes of research evaluation (Waltman *et al*., 2012; Waltman, Van Eck, Van Leeuwen, Visser, & Van Raan, 2011). The focus is no longer on impact as an attribute of the journal but on the journal as an information production process (Egghe & Rousseau, 1990). The relevant indicators should accordingly be appropriately sophisticated. In this paper, we propose to use four percentile rank classes (top-1%, top-10%, top-50%, and bottom-50%) as a categorical structure. These categories must be weighted for integration.

Our basic assertion is that a paper in the top-1% class is weighted at ten times the value of any other paper in the top-10% class. It follows then log-linearly that a top-1% paper weights 100 times more than a paper at the bottom. This weighting scheme appreciates the highly-skewed nature of citation distributions. We add, as a second assertion, a weighting to distinguish between papers in the top-50% and bottom-50% by weighting the latter with one and the former with two. The dividing line between bottom-50% and top-50% is less pronounced than the line between an averagely-cited paper and an exceptionally-cited one.

Table 1: *PLOS One* data as an example of the calculation of *I3*, based on non-normalized and field-normalized values

	Data from the WoS database		***Distinct classes***	***Number of papers in distinct classes***		***Weights***	***I3 and I3F***	
	Non-normalized (a)	*Field-normalized (b)*	*Percentile rank classes*	*Non-normalized (c)*	*Field-normalized (d)*		*Non-normalized (f)*	*Field-normalized (g)*
Top 1%	91	14.000	99-100	91	14.000	x 100 =	9100	1400
Top 10%	2,545	926.821	90-98	2,454	912.821	x 10 =	24,500	9,128.21
Top 50%	20,141	14,853.688	50-89	17,506	13,926.867	x 2 =	35,192	27,853.73
			0-49	9,901	15,188.312	x 1 =	9,901	15,188.31
Total	30,042	30,042		30,042	30,042		**78,733**	**53,570.26**

Table 1 shows how to calculate *I3* based on publication numbers, using *PLOS One* as an example. The publication numbers in the first columns (a and b) are obtained from Clarivate's *Web of Science* (WoS) database. These are the numbers of papers in the different top-*x*%-classes. Since the publication numbers in the higher classes are subsets of the numbers in the lower classes, the percentile classes are corrected (by subtraction) to avoid double counting. The resulting values in each distinct class are provided in the columns c and d. The distinct class counts are then multiplied by the appropriate weights. In the last step of calculating *I3*, the weighted numbers of papers in the distinct classes are summed into *I3.* In this case, *I3* = 78,733 (non-normalized) and *I3F* = 53,570.26 (field-normalized). The template available at https://www.leydesdorff.net/*I3*/template.xlsx automatically fills out the numbers and significance levels when the user provides the field-normalized and non-normalized values for top-1%, top-10%, top-50%, and total number of papers in the respective cells.

Methodology

Data were harvested in collaboration with the Max Planck Digital Library (MPDL) at the in-house database of the Max Planck Society during the period October 15-29, 2018. This database contains an analytically enriched copy of the *Sciences Citation Index*-Expanded (SCI-E), the *Social Sciences Citation Index* (SSCI), and the *Arts and Humanities Citation Index* (AHCI). Citation count data can be normalized for the Clarivate *Web of Science* Subject Categories (WoSCATs)

and theoretically could be based on whole-number counting or fractional counting in the case of more than a single co-author. The citation window in the in-house database is currently until the end of 2017. We first collected data for substantial items (articles and reviews) using the publication year 2014 with a three-year citation window until the end of 2017. Thereafter, the results were checked against a similar download for the publication year 2009, that is, five years earlier. The year 2014 was chosen as the last year with a complete three-year citation window at the time of this research (October-November, 2018); the year 2009 was chosen because it is the first year after the update of WoS to its current version 5.

Citation counts can be field-normalized using the WosCATs. These field-normalized scores are available at individual document level for all publications since 1980 in the in-house database. The *I3* index calculated with field-normalized data is denoted as *I3F* for clarification. When journals are assigned to more than one WoSCAT, the journal items and their citation counts are fractionally attributed. In the case of ties at the thresholds of a top-*x*% class of papers (see above), the field-normalized indicators have been calculated following Waltman & Schreiber (2013). Papers at the threshold separating the top from the bottom are fractionally assigned to the top paper set.

As noted, *I3* can be divided by *N*, the number of publications (which is by definition equal to the sum of the numbers in the four percentile classes). *I3/N* is based on relative frequencies, since the number in each term (n_i) is divided by N ($= \Sigma_i\ n_i$). One can expect *I3/N* to no longer be size-dependent and thus to have different applications from *I3*, as we shall show below. We focus on *I3* in this paper; we will discuss potential applications of *I3/N* in a later paper.

We have applied Spearman rank-correlation analysis and factor analysis (Principal Component Analysis with varimax rotation) to the following variables:

1. total numbers of publications (*NPub*);
2. citations (*NCit*);
3. *JIF*2;
4. *JIF*5;
5. Non-normalized *I3*-values (*I3*);
6. Field-normalized *I3*-values (*I3F*);
7. *I3/N* for the non-normalized case (*I3/N*).

The results are shown as factor-plots using the first two components as *x*- and *y*-axes. This representation provides a ready means of assessing the results visually. Rotated factor matrices and the percentages of explained variance will also be provided for each analysis. The results indicate that the two first eigenvalues explain about 85-90% of the variance in the subsequent analyses.

Results

Full set (journal count, n = 10,942)

Figure 1 shows the two-dimensional factor plot of the factor loadings provided numerically in Table 2. These first two factors explain 87.5% of the variance. The correlation between *I3* and its field-normalized equivalent *I3F* and the first component is greater than 0.9, so they can be considered as essentially the same characteristic. The factor loadings of the numbers of citations

(*NCit*) and publications (*NPub*) on this first factor are greater than 0.8. *NPub*, which is the size indicator of output (number of publications), does not load substantially on the second factor which represents impact (number of citations); however, the number of citations loads on factor 1 (.802) much more than on factor 2 (.304).

Table 2: Rotated factor matrix of the seven indicators plotted in Figure 1.

	Component	
Indicator	1	2
I3F	**.925**	.284
I3	**.915**	.286
NPub	**.870**	-.094
NCit	**.802**	.304
*JIF*5	.175	**.958**
*JIF*2	.188	**.957**
I3/N	.162	**.917**

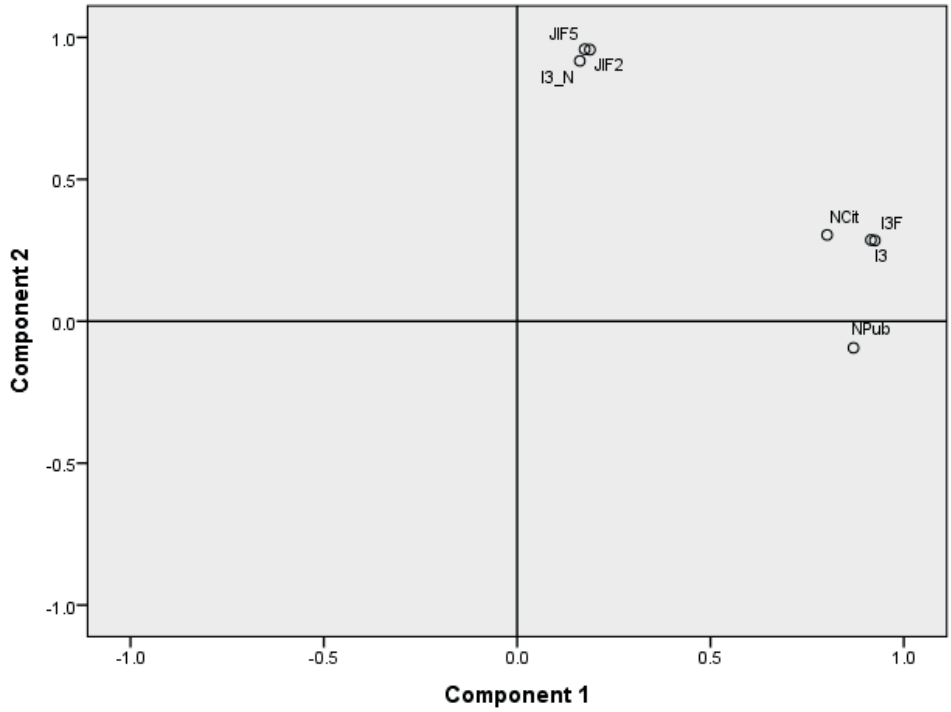


Figure 1: Component plot in rotated space of the two main components in the matrix (varimax-rotated PCA) of 10,942 cases (journals) versus seven indicators.

Notes: The seven indicators are: total numbers of publications (*NPub*); citations (*NCit*); *JIF*2; *JIF*5; non-normalized *I3*-values (*I3*); field-normalized *I3*-values (*I3F*); and scaled *I3* for the non-normalized case (*I3*_N).

Table 3: 25 journals ranked on non-normalized *I3* values (*I3*), field-normalized values (*I3F*), and *I3/N* (non-normalized). For the full list see http://www.leydesdorff.net/*I3*/ranking.htm .[1]

JOURNAL	*I3*	*JOURNAL*	*I3F*	*JOURNAL*	*I3 / N*
PLOS One	78,733	*PLOS One*	53,570.26	Nat. Rev. Drug Discov.	90.8
J. Am. Chem. Soc.	55,786	Nature	27,397.23	Physiol. Rev.	76.8
Nature	52,888	Phys. Rev. Lett.	24,909.74	Nat. Rev. Genet.	72.5
Proc. Natl. Acad. Sci. U. S. A.	47,041	Adv. Mater.	23,741.25	Prog. Mater. Sci.	71.9
Nat. Commun.	46,762	Nat. Commun.	21,689.35	Nat. Rev. Mol. Cell Biol.	70.6
Science	41,946	Science	21,493.38	Nat. Rev. Cancer	69.8
Angew. Chem.-Int. Edit.	40,572	Proc. Natl. Acad. Sci. U. S. A.	21,204.46	Chem. Rev.	63.7
Adv. Mater.	34,435	J. Am. Chem. Soc.	20,435.29	Nat. Rev. Neurosci.	63.1
Phys. Rev. Lett.	30,549	J. Mater. Chem. A	18,323.90	Nat. Rev. Immunol.	62.1
ACS Nano	29,284	Nano Lett.	16,905.86	N. Engl. J. Med.	62.0
J. Mater. Chem. A	28,260	ACS Nano	16,608.79	Nature	61.4
Chem. Commun.	26,209	Phys. Rev. B	16,176.17	Living Rev. Relativ.	57.7
Nano Lett.	24,717	Appl. Phys. Lett.	16,077.51	Chem. Soc. Rev.	56.8
ACS Appl. Mater. Interfaces	24,407	Angew. Chem.-Int. Edit.	15,136.02	Lancet	55.8
RSC Adv.	24,326	Opt. Express	14,893.89	Rev. Mod. Phys.	55.7
Cell	22,993	Org. Lett.	14,819.56	Cell Stem Cell	54.6
N. Engl. J. Med.	21,874	Energy Environ. Sci.	14,565.50	Nat. Photonics	54.4
Chem. Soc. Rev.	21,576	RSC Adv.	14,479.96	Nature Genet.	54.2
Sci Rep	20,098	Cell	13,689.86	Prog. Polym. Sci.	53.5
Nanoscale	19,631	ACS Appl. Mater. Interfaces	13,678.24	Nat. Biotechnol.	53.2
Astrophys. J.	19,130	Anal. Chem.	12,548.23	Psychol. Sci. Public Interest	52.8
Phys. Rev. B	18,831	Phys. Chem. Chem. Phys.	12,459.76	Cell	52.7
Chem. Rev.	17,889	Nanoscale	12,418.65	Nat. Med.	52.5
Energy Environ. Sci.	17,151	Astrophys. J.	12,269.03	Nat. Mater.	51.3
Phys. Rev. D	16,893	Chem. Soc. Rev.	11,710.45	Cancer Cell	51.2

[1] For the sake of readability, journal names in this table follow the ISO abbreviations instead of those of the ISI.

Table 3 shows the ranking of the 25 journals with the greatest index values for each of *I3,* field-normalized *I3F,* and *I3/N*, respectively. (The full listing is available at http://www.leydesdorff.net/*I3*/ranking.htm.) The size effect of *PLOS One* dominates both the *I3* and *I3F* ranking, but not the third column (*I3/N*) which is size-independent because of the division by volume. Twelve of the 25 titles in this latter column are attributed to *Nature* or specialist journals in the *Nature* publishing group. Note that *Science*, which occupies sixth position in the first two columns, drops to 29th position on the size-independent indicator. *PLOS One* falls much further, to position 2,064. Among journals ranked by non-normalized *I3*, there are many journals belonging to the life sciences, and only a few to physics or chemistry. Among the ranking based on field-normalized *I3*, however, there are only a few journals from the the life sciences.

The Social Sciences Citation Index (SSCI)

The citation environment of journals listed in the *Social Sciences Citation Index* is very different from that of journals in the SCI-E. The SCI-E journals in JCR constitute about 28% (3,105 / 10,942) of the total serial titles, but the total citations to SSCI journals constitute less than 10% of all citations to JCR titles (4,506,510/48,340,046 in our time window). The average yearly total cites (*NCit*) of a journal in SSCI is 1,451.3 compared with 4,417.8 for the combined set.

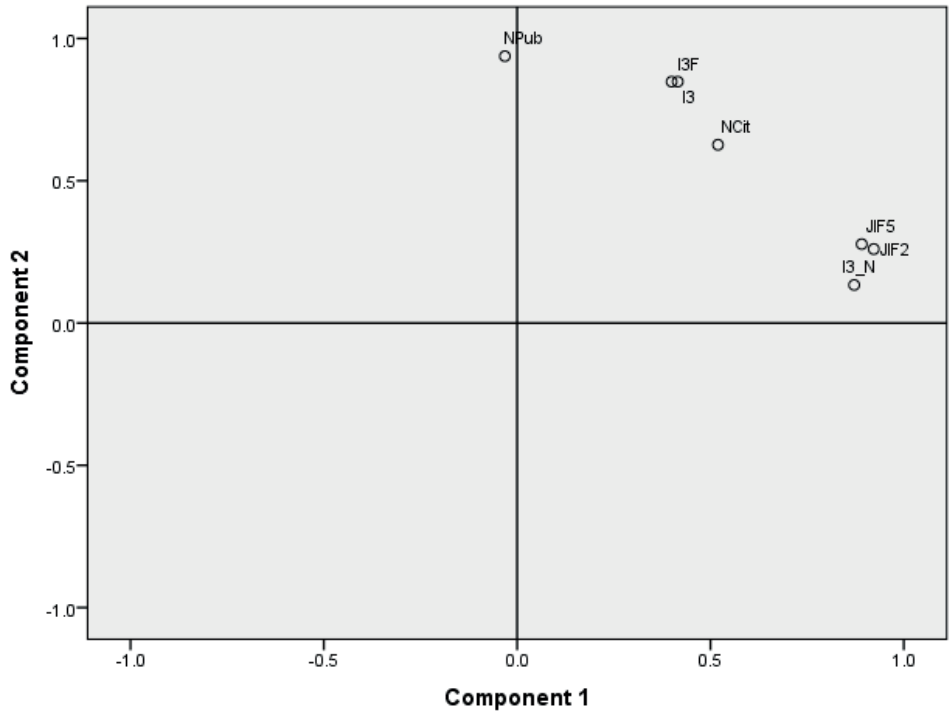


Figure 2: Two-component factor plot of seven indicators for 3,105 journals in the SSCI.
The indicators are: total numbers of publications (*NPub*); citations (*NCit*); *JIF*2; *JIF*5; non-normalized *I3*-values (*I3*); field-normalized *I3*-values (*I3F*); and scaled *I3* for the non-normalized case (*I3*_N).

Table 4: Rotated factor matrices for journals in the Social Science Citation Index and for journals in two specialist WoSCATs, one each from SSCI and SCI.

Social Science Citation Index (3,105 journals)

Rotated Component Matrix[a]

Indicator	Component	
	1	2
*JIF*2	**0.922**	0.26
*JIF*5	**0.891**	0.277
I3/N	**0.872**	0.134
NPub		**0.938**
I3	0.415	**0.848**
I3F	0.399	**0.848**
NCit	***0.519***	***0.626***

Library and Information Science (83 journals)

Rotated Component Matrix[a]

Indicator	Component	
	1	2
*JIF*5	**0.959**	0.170
*JIF*2	**0.927**	0.215
I3/N	**0.799**	0.265
NCit	***0.757***	***0.527***
NPub		**0.968**
I3	0.41	**0.903**
I3F	0.444	**0.861**

Spectroscopy (41 journals)

Rotated Component Matrix[a]

Indicator	Component	
	1	2
I3	**0.982**	
I3F	**0.962**	
NPub	**0.959**	-0.113
NCit	**0.771**	0.167
*JIF*2		**0.987**
*JIF*5		**0.977**
I3/N		**0.930**

Notes: Extraction Method: Principal Component Analysis. Rotation Method: Varimax with Kaiser Normalization. The indicators are: total numbers of publications (*NPub*); citations (*NCit*); *JIF*2; *JIF*5; non-normalized *I3*-values (*I3*); field-normalized *I3*-values (*I3F*); and scaled *I3* for the non-normalized case (*I3/N*).

[a.] Rotation converged in 3 iterations.

Figure 2 shows the factor plot for the 3,105 journals in SSCI for comparison with the factor plot for the combined sets of SSCI and SCI provided in Figure 1. The main difference is the greater distance between the number of citations (*NCit*) and the number of publications (*NPub*). The correlation between both indicators is smaller in SSCI (.633) than in SCI (.719). Consequently, *NPub* and *NCit* are distanced in Figure 2 and the order of the two factors is reversed. Nonetheless, these two factors together still explain 84% of the variance.

If we focus on a specific journal category of SSCI, such as the 83 journals in Library & Information Science, the difference between SCI and SSCI outcomes is further emphasized. Alternatively, if we focus on a narrow specialism in the natural sciences, such as Spectroscopy with 41 journals, we find that the distinction between the two components is even more pronounced than for the full set of 10,942 journals. Table 4 juxtaposes the rotated factor matrices showing these differences numerically. While the number of publications drives the number of citations in the SCI-E, this appears to be less the case in the SSCI. Size is less important for impact in SSCI than in SCI-E. *I3* correlates with size (*NPub*) more than with citations (*NCit*) in the social sciences.

Comparison with 2009

It is possible that the results obtained for 2014 were specific for the year 2014 because it is recent and the citation counts were not yet stable. We tested this by repeating the analysis for 2009 data, which was chosen because the Web of Science (version 5) was reorganized in 2008/2009. Of the 9,216 journals in the combined sets of JCRs for SCI-E and SSCI, 8,994 journal title abbreviations could automatically be matched between the data from the in-house database and JCR.

The outcome for 2009 data is very similar to that seen with 2014 data (Table 5). Two factors explain 88.1% of the variance in 2009 and 87.5% in 2014. The results are virtually identical in these two sample years. Thus, the indicator appears to be robust over time.

Table 5: Rotated factor matrices for full sets in 2009 and 2014.

JCR 2009: 8,904 journals			**JCR 2014: 10,942 journals**		
	Component			Component	
	1	2		1	2
NPub	**0.904**	-0.048	***I3F***	**0.925**	0.284
I3F	**0.903**	0.276	***I3***	**0.915**	0.286
I3	**0.884**	0.329	***NPub***	**0.870**	
NCit	*0.87*	*0.259*	*NCit*	*0.802*	*0.304*
JIF5	0.212	**0.949**	*JIF5*	0.175	**0.958**
JIF2	0.201	**0.939**	*JIF2*	0.188	**0.957**
I3/N	0.145	**0.923**	***I3/N***	0.162	**0.917**

Statistics

PLOS One was by far the largest journal in 2014 with 30,042 publications. It was followed in this analysis by *RSC Advances* with 8,345 citable items. In terms of total citations, however, *PLOS One* is in eighth place with 332,716 citations. In the same year, *Nature* accrued 617,363 citations to 862 publications. The citations/publication (c/p) ratio for *Nature* is 716.3 and for *PLOS One* is 11.1. By comparison, the values of *I3/N* are 61.4 for *Nature* and 2.6 for *PLOS One* and, in seeming contradiction to conventional indicators, the (non-normalized) *I3* values are 78,733 for *PLOS One*

and 52,883 for *Nature.* Are the differences statistically and practically significant? One can test the distribution of papers over the classes against the expected numbers or against each other. This can be done for the frequencies in the matrix using chi-square statistics, or by a test between means (in the case of *I3/N*) using the *z*-test and/or Cohen's **h** index for "practical significance."[1] Table 5 shows various options for testing *PLOS One* versus *RSC Advances*.

The results of the chi-square tests are statistically significant ($p < .001$), when comparing *PLOS PLOS One* with *RSC Advances.* One can summarize the results of the chi-square *ex post* using Cramèr's *V* which conveniently ranges from zero to one. In this case, Cramèr's $V = 0.05$. In other words, the differences between the expected and observed percentile-rank distribution is more than five times larger than the corresponding differences between *PLOS One* and *RSC Advances.* The results of the chi-square statistic based on testing the *I3* values (in columns g and h in) are provided in column k at the bottom.

While the chi-square statistic provides a test for comparing the entire distributions (two vectors of four classes), the decomposition of chi-square into standardized residuals provides us with a statistic for each class. Standardized residuals can be considered as *z*-values: they are statistically significant at the 5% level if the absolute value is larger than 1.96, 1% for an absolute value > 2.576, and 1‰ for an absolute vale > 3.291 (Sheskin, 2011, at p. 672).

Furthermore, the residuals are signed and indicate in Table 6 that *RSC Advances* scores statistically significantly higher than *PLOS One* in the top-10% (column l), but not statistically-significant below *PLOS One* in the lower-ranked classes. Tables 6b add the statistics for *I3/N.* The division by *N* makes all the frequencies relative. Since these relative frequencies can also be considered as proportions, one can *z*-test for difference in proportions (Sheskin, 2011, pp. 656f.) or also compute an effect size using Cohen's **w** (1988, at p. 216; Leydesdorff, Bornmann, & Mingers, 2018, forthcoming).

These results may come as no surprise, but in cases other than *PLOS One* may offer less intuitive results about the status of a journal. For example, specification of the differences between *RSC Advances* and *Nature* in terms of these four classes would be far from obvious. The template available at https://www.leydesdorff.net/*I3*/template.xlsx automatically fills out the numbers and significance levels when the user provides the field-normalized and non-normalized values for top-1%, top-10%, top-50%, and total number of papers in the respective cells.

[1] Cohen's **h** tests proportions against each other for each row using $\mathbf{h} = 2^* (\arcsin\sqrt{p_{obs}} - \arcsin\sqrt{p_{exp}})$ (Cohen, 1988, pp. 180 ff.), whereas Cohen's **w** first sums over the rows and then takes the square root (Cohen, 1988, pp. 216f.): $\mathbf{w} = \sqrt{\sum_{i=1}^{m} \frac{(p(observed)-p(expected)\text{^}2}{p(expected)}}$.

Table 6: Comparison of *PLOS One* (unit 1) with *RSC Advances* (unit 2)

Class (a)	*unit 1* (b)	*unit2* (c)	*Class* (d)	*n1* (e)	*n2* (f)	*I3_1* (g)	*I3_2* (h)	*standardized residuals of the χ2* (i)	(j)	*χ2* (k)
top-1%	91	30	**99-100**	91	30	9100	3000	1.196	*n.s.*	9.493
top10%	2545	909	**90-98**	2454	879	24540	8790	4.621	*p <.001*	141.686
top-50%	20141	5919	**50-89**	17596	5010	35192	10020	-2.802	*p < .01*	52.113
bottom-50%	7265	1577	**0-49**	9901	2516	9901	2516	-3.404	*p <.001*	76.881
Sum	**30042**	**8435**		**30042**	**8435**	**78733**	**24326**		***χ2 =***	**280.173**
									df = 3	***p<.001***
									Cramèr's V = 0.0521	

Class (a)	*I3/N unit1* (l)	*I3/N unit2* (m)	*p1* (o)	*p2* (p)	*z-test* (q)	(r)	*Cohen's w* (s)	*Cohen's h*
top-1%	0.303	0.356	0.003	0.004	0.765	*n.s.*	0.000	0.009
top10%	0.817	1.042	0.082	0.104	6.498	*p<.001*	0.005	0.078
top-50%	1.171	1.188	0.586	0.594	1.358	*n.s.*	0.000	0.017
bottom-50%	0.330	0.298	0.330	0.298	-5.432	*p<.001*	0.003	-0.067
Sum	2.621	2.884	**1.000**	**1.000**			**0.091**	

* critical values		**χ2; df = 3**
	***p* < 0.001**	16.266
	***p* < 0.01**	11.345
	***p* < 0.05**	7.815

z

In order to have information about the significance of the results on the basis of effect sizes (Cohen, 1988; Schneider, 2013; Wasserstein & Lazar, 2016; Williams & Bornmann, 2014), we added Cohen's **h** and **w** for the comparison among proportions as columns to Table 6. The **w** index is 0.091 in Table 6, and thus the difference between the two journals is not meaningful for practical purposes. However, these tests on proportions address the size-independent indicator *I3/N*. This measure can be used as the expected value of citations of a publication published in the relevant journal: a paper that is accepted for publication in *RSC Advances* has a significantly greater likelihood of being cited in the overall top-10% than a paper in *PLOS One*. It is also less likely to be cited below the 50% threshold.

Summary and conclusions

We argue in this paper that an indicator can be developed that reflects both impact and performance, and that combines the two dimensions of publications and citations into a single measure by using non-parametric statistics. This indicator (*I3*) can be size-normalized by dividing by the number of publications to obtain a secondary indicator of the expected contribution made by a single paper given the journal's characteristics. The size-normalized and size-independent indicator can be considered as relating to two nearly orthogonal axes. When we consider the relationship between conventional journal indicators and these new integrated indicators, we see that *I3* correlates strongly with both the total number of citations and publications, whereas *I3/N* correlates with size-independent indicators such as *JIF*.

The versatility of *I3* is illustrated in a spreadsheet in Excel containing a template for the computation. The $P_{\text{top 10\%}}$ and $\text{PP}_{\text{top 10\%}}$ indicators have become established as quasi-standard indicators in professional bibliometrics, especially when research institutions are compared (Waltman *et al*., 2012). The use of these percentile-based indicators is recommended, for instance in the Leiden Manifesto, which included ten guiding principles for research evaluation (Hicks *et al*., 2015). It is an advantage of the *I3* indicator—which is a percentile-based indicator—that it integrates the top-1% with the top-10% information and combines them with information about other percentile classes. Thus, one provides a broader picture by using *I3* as indicator compared to $P_{\text{top 10\%}}$ and $\text{PP}_{\text{top 10\%}}$. One can consider as a disadvantage of percentile-based indicators that they can only be computed with reference to a set, and are therefore not characteristics of individual papers.

Acknowledgements

The bibliometric data used in this paper are from an in-house database developed and maintained in collaboration with the Max Planck Digital Library (MPDL, Munich) of the Max Planck Society, and derived from the Science Citation Index Expanded (SCI-E), the Social Sciences Citation Index (SSCI), and the Arts and Humanities Citation Index (AHCI) prepared by ISI/ Clarivate Analytics (Philadelphia, Pennsylvania, USA). The journal data are based on the JCR provided by ISI/ Clarivate Analytics.

References

Bornmann, L., & Mutz, R. (2011). Further steps towards an ideal method of measuring citation performance: The avoidance of citation (ratio) averages in field-normalization. *Journal of Informetrics, 5*(1), 228-230.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences (2nd ed.)*. Hillsdale, NJ: Lawrence Erlbaum.

Egghe, L., & Rousseau, R. (1990). *Introduction to Informetrics*. Amsterdam: Elsevier.

Garfield, E. (1972). Citation Analysis as a Tool in Journal Evaluation. *Science 178*(Number 4060), 471-479.

Garfield, E. (1979). Is citation analysis a legitimate evaluation tool? *Scientometrics, 1*(4), 359-375.

Garfield, E. (2003). The Meaning of the Impact Factor. *Revista Internacional de Psicologia Clinica y de la Salud, 3*(2), 363-369.

Garfield, E. (2006). The history and meaning of the journal impact factor. *JAMA, 295*(1), 90-93.
Garfield, E., & Sher, I. H. (1963). New factors in the evaluation of scientific literature through citation indexing. *American Documentation, 14*(3), 195-201.
Hicks, D., Wouters, P., Waltman, L., de Rijcke, S., & Rafols, I. (2015). Bibliometrics: The Leiden Manifesto for research metrics. *Nature, 520*(7548), 429-431.
Leydesdorff, L., & Bornmann, L. (2011). Integrated Impact Indicators Compared With Impact Factors: An Alternative Research Design With Policy Implications. [Article]. *Journal of the American Society for Information Science and Technology, 62*(11), 2133-2146. doi: 10.1002/asi.21609
Leydesdorff, L., Bornmann, L., Mutz, R., & Opthof, T. (2011). Turning the Tables on Citation Analysis One More Time: Principles for Comparing Sets of Documents. [Article]. *Journal of the American Society for Information Science and Technology, 62*(7), 1370-1381. doi: 10.1002/asi.21534
McAllister, P. R., Narin, F., & Corrigan, J. G. (1983). Programmatic Evaluation and Comparison Based on Standardized Citation Scores. *IEEE Transactions on Engineering Management, 30*(4), 205-211.
Narin, F. (1987). Bibliometric techniques in the evaluation of research programs. *Science and Public Policy, 14*(2), 99-106.
Pendlebury, D. A. & Adams, J. (2012). Comments on a critique of the Thomson Reuters journal impact factor. *Scientometrics*, *92*, 395-401. doi: 10.1007/s11192-012-0689-6
Schneider, J. W. (2013). Caveats for using statistical significance tests in research assessments. *Journal of Informetrics, 7*(1), 50-62.
Seglen, P. O. (1992). The Skewness of Science. *Journal of the American Society for Information Science, 43*(9), 628-638.
Seglen, P. O. (1997). Why the impact factor of journals should not be used for evaluating research. *British Medical Journal, 314*, 498-502.
Sheskin, D. J. (2011). *Handbook of Parametric and Nonparametric Statistical Procedures (5th Edition)*. Boca Raton, FL: Chapman & Hall/CRC.
Tijssen, R. J. W., Visser, M. S., & Van Leeuwen, T. N. (2002). Benchmarking international scientific excellence: are highly cited research papers an appropriate frame of reference? *Scientometrics, 54*(3), 381-397.
Waltman, L., & Van Eck, N. J. (2012). The inconsistency of the h-index. *Journal of the American Society for Information Science and Technology, 63*(2), 406-415.
Waltman, L., Van Eck, N. J., Van Leeuwen, T. N., Visser, M. S., & van Raan, A. F. J. (2011). Towards a New Crown Indicator: Some Theoretical Considerations. *Journal of Informetrics, 5*(1), 37-47.
Waltman, L., Calero-Medina, C., Kosten, J., Noyons, E., Tijssen, R. J., Eck, N. J., . . . Wouters, P. (2012). The Leiden Ranking 2011/2012: Data collection, indicators, and interpretation. *Journal of the American Society for Information Science and Technology, 63*(12), 2419-2432.
Waltman, L., & Schreiber, M. (2013). On the calculation of percentile-based bibliometric indicators. *Journal of the American Society for Information Science and Technology, 64*(2), 372-379.
Waltman, L., & Traag, V. A. (2017). Use of the journal impact factor for assessing individual articles need not be wrong. *arXiv preprint arXiv:1703.02334*.
Wasserstein, R. L., & Lazar, N. A. (2016). The ASA's statement on p-values: context, process, and purpose. *The American Statistician, 70*(2), 129-133.
Williams, R., & Bornmann, L. (2014). The substantive and practical significance of citation impact differences between institutions: Guidelines for the analysis of percentiles using effect sizes and confidence intervals. In Y. Ding, R. Rousseau & D. Wolfram (Eds.), *Measuring Scholarly Impact: Methods and practice* (pp. 259-281). Heidelberg: Springer.

Non-English language publications in Citation Indexes – quantity and quality.

Olga Moskaleva[1] and Mark Akoev [2]

[1] o.moskaleva@spbu.ru
Saint-Petersburg State University, Universitetskaya emb. 7-9, 199034, St. Petersburg, Russia

[2] m.a.akoev@urfu.ru
Ural Federal University, 19 Mira str., Ekaterinburg, 620002, Russia

Abstract.

We analyzed publications data in WoS and Scopus to compare publications in native languages vs publications in English and find any distinctive patterns. We analyzed their distribution by research areas, languages, type of access and citation patterns. The following trends were found: share of English publications increases over time; native-language publications are read and cited less than English-language outside the origin country; open access impact on views and citation is higher for native languages; journal ranking correlates with the share of English publications for multi-language journals. We conclude also that the role of non-English publications in research evaluation in non-English speaking countries is underestimated when research in social science and humanities is assessed only by publications in Web of Science and Scopus.

Introduction

The concept of English language as the modern lingua franca in science (Montgomery, 2016) is suggested by many studies. The number and proportion of English-language documents in Web of Science is growing (Kumar, Panwar, & Mahesh, 2016; W. Liu, 2017; Meneghini & Packer, 2007). Nevertheless, the total number of non-English publications is growing also as the number of researchers in non-English speaking countries (UNESCO, 2016) grows.

Available publications we have found focus mainly on the number of non-English language documents or the disadvantage which they receive in the number of citations (Di Bitetti & Ferreras, 2017; Fung, 2008) or journal impact factors (Diekhoff, Schlattmann, & Dewey, 2013; González-Alcaide, Valderrama-Zurián, & Aleixandre-Benavent, 2012; Liang, Rousseau, & Zhong, 2013; F. Liu, Hu, Tang, & Liu, 2018). Presumably, journal articles were analyzed.

The dominance of English language appears to be crucial at the second half of XX century, though in 1960-1970 Russian Language was the second in Natural Sciences (Ammon, 2012) after English, while at the beginning of XX English, French and German were almost in equal proportions. In Social Sciences, according to International Bibliography of the Social Sciences, French is the most popular scientific language together with English (Ammon, 2012). The same article underlines preferences of English language regardless of scientist's native languages – publications in English in almost all cases have a higher impact.

Michael D.Gordin (Gordin, 2015) gives a brilliant review of the role of different languages in science in a monograph. He describes reasons and consequences of English language dominance in scientific literature, differences in natural and social sciences, etc.

There are more than seven thousands of languages in the world (though not all of them have a written form). The most widely spread language by the number of speakers is Chinese (1,2 billion), the second is Spanish, the third – English, but the number of countries where English is the main language is three times higher than Spanish – 106 English-speaking countries against 31 for Spanish (Ethnologue, 2015).

Non-English language publications in citation indexes

The number of non-English documents in Web of Science or Scopus does not correlate neither with number of speakers, nor with the number of countries. We can see no correlation with the number of researchers with corresponding native language either. In 2013 there were 7,8 million researchers, among them 25% with English being native or one of state languages, Chinese - 17%, Japanese - 8,5%, Russian - 6%, German - 5%, Spanish - 3% and Portugal - 2% (UNESCO, 2016).

On the other hand, many research publications are not indexed in global citation databases, such as Web of Science and Scopus, especially in Social Science and Humanities (Kulczycki et al., 2018). The study of publication output in these areas in 7 European non-English speaking countries demonstrated that only part of them are covered by Web of Science (50,4% for Denmark and 15% for Poland), so research evaluation may be not correct when using only Web of Science data. This obstacle is the serious reason for national citation databases emerging globally (Pislyakov, 2007), enabling more adequate research assessment, especially in Social science & Humanities (Sīle, Guns, Sivertsen, & Engels, 2017). Some of independently created national databases become basis for national citation indexes on Web of Science platform, i.e. Chinese Citation Index, SciELO, Korean Citation Index and Russian Science Citation Index (Moskaleva, Pislyakov, Sterligov, Akoev, & Shabanova, 2018). By the end of May 2018 Clarivate Analytics announced plans for creating of a new citation index: Arabic Citation Index (ARCI) is to be launched in 2020 ("Clarivate Analytics Partners with the Egyptian Knowledge Bank to Power the First Arabic Citation Index - Clarivate," 2018).

In contrast to the opinion widely spread in non-English countries, global citation indexes include documents not only in English, and thanks to the document indexing policies, only the best material (journals, conference proceedings, books, etc.) is presented in Web of Science and Scopus. This obstacle makes it possible to assess the national science in total and the role of national languages in global science. A lot of information about publications in national languages can be extracted from reference lists in indexed documents, so everybody can draw their own opinion about national publications, cited in thoroughly selected sources.

The comparison of language-speaking countries and countries with publications in certain languages demonstrates more language diversity for research communication (Fig.1). Certainly, mainly it happens due to international collaboration, the main language of which is English (not shown on diagram).

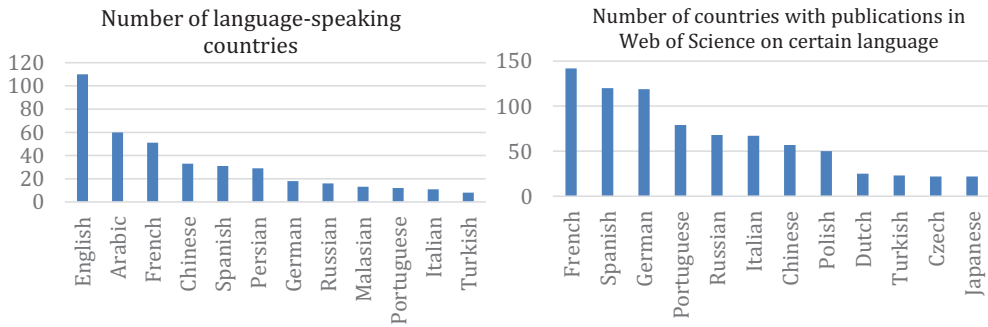


Fig.1. Ranking of language-speaking countries and affiliated countries in Web of Science publications (2007-2016).

The total number of non-English publications is growing from the beginning of XXI century both in Web of Science CC and in Scopus, though their share in total output diminishes (Fig.2). The proportion of English and non-English publications differs in various indexes in Web of Science CC and different publication types in Scopus. Less representation of non-English language journals and, as a consequence, of non-Anglophone publications in Web of Science CC as to compare with Scopus is compensated by development of national citation indexes on Web of Science platform. Such indexes are fully integrated with all other databases on Web of Science platform (Core Collection, BIOSIS, Medline, Derwent, etc.), so researchers in the whole world can have access to non-English language science of a high quality. Using Web of Science or Scopus as a search engine for publications provides researchers content of a higher quality in comparison with Google Scholar, for example, because of thoroughly selected sources and vast range of refine possibilities. All these tools help researcher to find relevant content with less efforts and time independently of original language of publication.

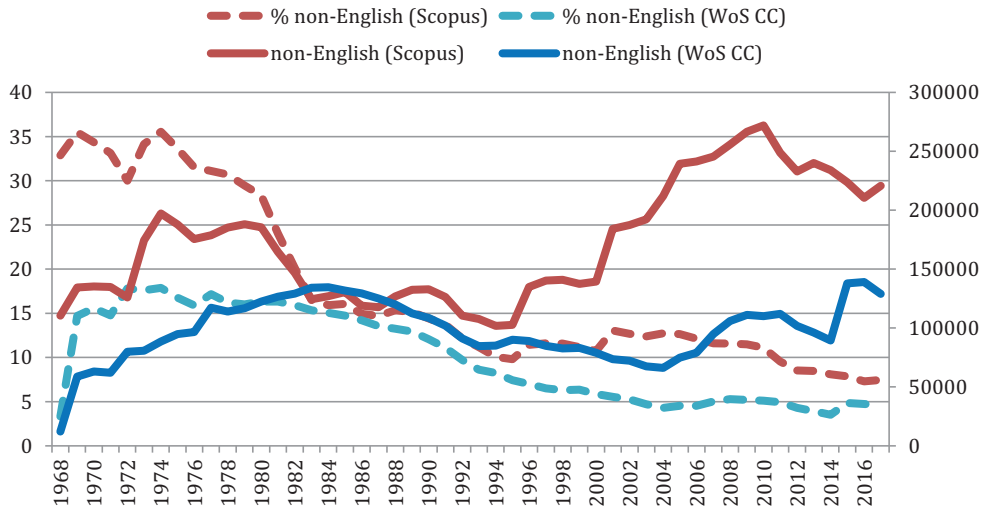


Fig.2 The total number and proportion of non-English documents in Web of Science CC and Scopus

Thus the international language in science now remains English, as earlier it was Latin, French or German. The growing publication output and proportion of English-language documents in global citation databases approves this fact. The common language for science makes an easier communication between researchers from different countries possible due to the common terminology. It allows attracting more scientists to the development of scientific knowledge, and on the other hand allows scientists with fluent English to have access to all scientific knowledge, and not just to a subset that is translated into national languages.

There are significant differences between research areas with respect to the role of non-English language documents. In Web of Science Core Collection for example, the share of non-English language publications in Arts & Humanities Citation Index exceeds 25% of total research output. Relatively high proportion of non-English documents is a characteristic feature of Emerging Sources Citation Index. Conference Proceedings in Social Science & Humanities (CPCI-SSH) also demonstrate 7 times higher proportion of non-English documents than in Natural Sciences and

Engineering (CPCI-S). The same holds true for Book Citation Indexes (Fig.3). The dynamics of non-English documents in citation indexes in Web of Science CC also differs significantly (W. Liu, 2017).

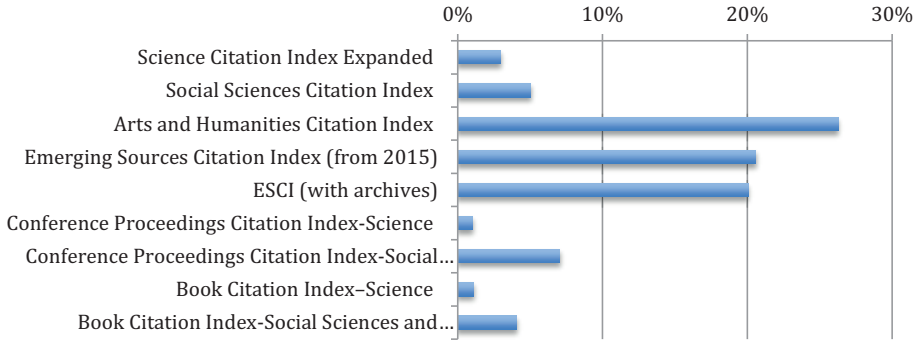


Fig.3 Distribution of non-English language documents among citation indexes in Web of Science CC (2007-2016, % of total number of documents).

In Scopus we can see differences in dynamic of the Non-English publication share for different source types. It is decreasing for journals and relatively stable for books and conference proceedings (Fig.4).

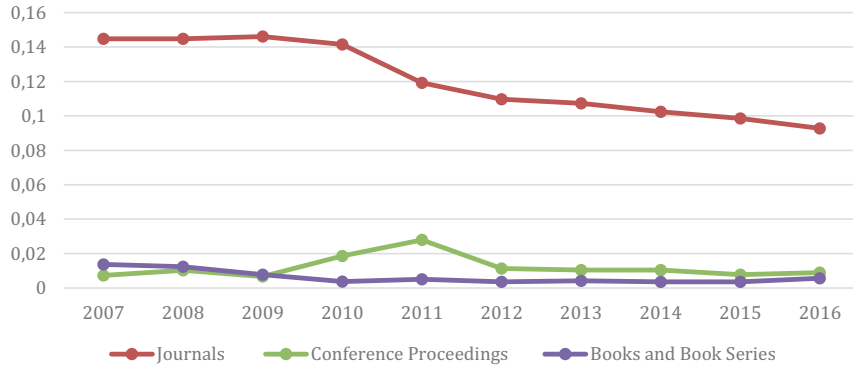


Fig.4 The proportion of non-English language documents in different source types in Scopus

The distribution of non-English documents by languages in Web of Science and Scopus looks very similar except for Chinese leading in Scopus and being only on the fourth place in Web of Science (Fig.5).

Even here we can see differences among different citation indexes in Web of Science Core Collection – German dominates in SCI-E, SSCI, BKCI-S and BKCI-SSH, the main language after English in AHCI is French, in CPCI-S and CPCI-SSH the main language is Chinese. In ESCI the highest share of non-English documents belongs to documents in Spanish.

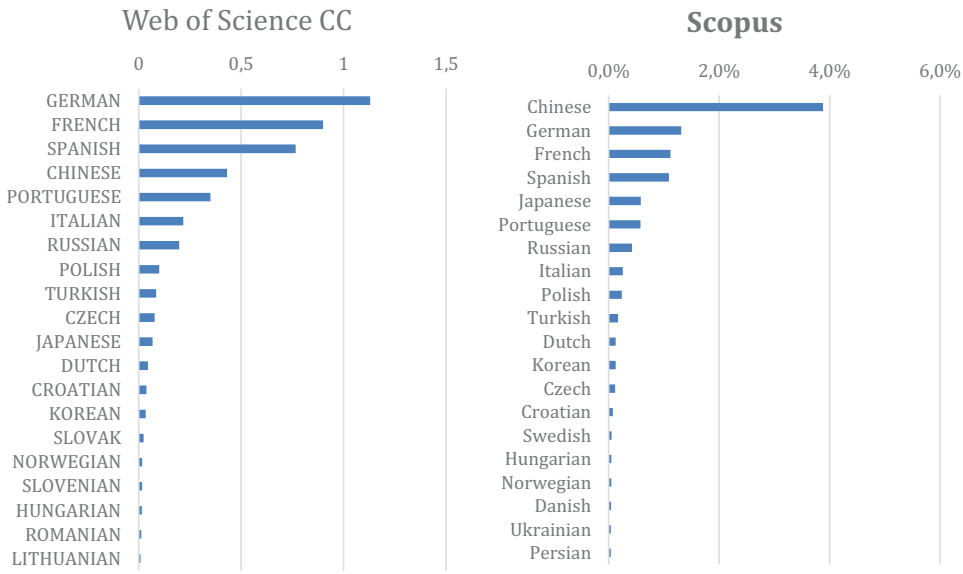


Fig.5 The language distribution of non-English language documents in Web of Science CC and Scopus (2007-2016, % of total number of documents).

Large body of research demonstrates the growing role of open access, which now influences both visibility and citedness of scientific results. New options in citation indexes were launched in 2017 – beginning of 2018 that allow finding articles available in gold (both in Web of Science and in Scopus) and greening (only in Web of Science) open access. These options make it easier to analyze their role in science development. The "elder" journal indexes of Web of Science CC (SCI-E, SSCI, AHCI) include up to 25% of open access documents (in SCI-E), less in CPCI and BKCI. Among non-English language publications, the best proportion of OA is observed for SCI-E – over 15%. ESCI is the most "open" citation database in Web of Science CC and includes 35% of OA documents, with more than 27% of non-English language publications available in open access (Fig.6).

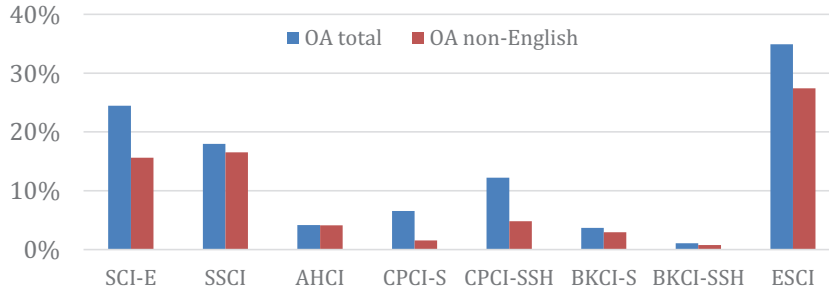


Fig.6 The proportion of Gold open access publications in Web of Science CC (2007-2016)

The languages distribution of OA publications by the share of OA documents in the total output in the same language is demonstrated on Fig.7.

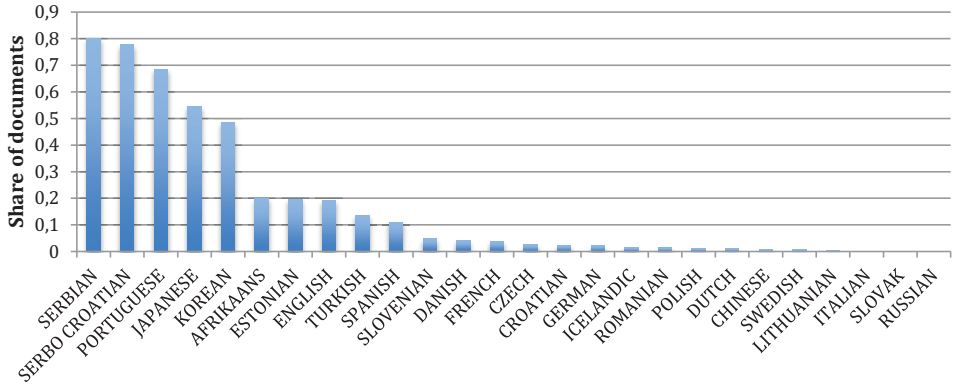


Fig.7 The proportion of Gold OA in journal papers in different languages (2006-2017, SCI-E, SSCI, AHCI)

Serbian, Portuguese, Japanese and Korean languages appeared to be the most open among other non-English papers in journals of "elder" journal indexes of Web of Science CC with 50-80% of Gold OA documents.

To determine the citation impact of open access publications we created corresponding data sets and analyzed them in InCites – analytical instrument of Clarivate Analytics. It provides a lot of complex bibliometric indicators and allows comparing different publication datasets independently of publication date, source and subject category. The results on Fig.8 show that both average citation and CNCI are almost twice higher for OA publications in Portuguese and Japanese than for traditional publications that are available only with subscription. At the same time, there are almost no non-English language publications among highly cited documents, independently of access type. In total, their citation is much lower than of English-language documents.

Researchers from different countries use their national languages to present their results to a different extent. The proportion of documents in Portuguese in Brazil is almost 14%, while in Portugal only 3% of total number of documents in Web of Science CC are in Portuguese (Fig.9).

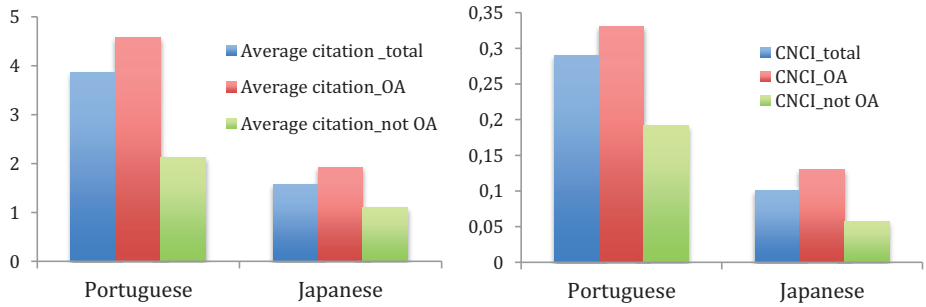


Fig.8 Average citation and Category Normalized citation impact (CNCI) of journal articles in Japanese and Portuguese languages with different access types (Web of Science, 2008-2016, SCI-E, SSCI, AHCI)

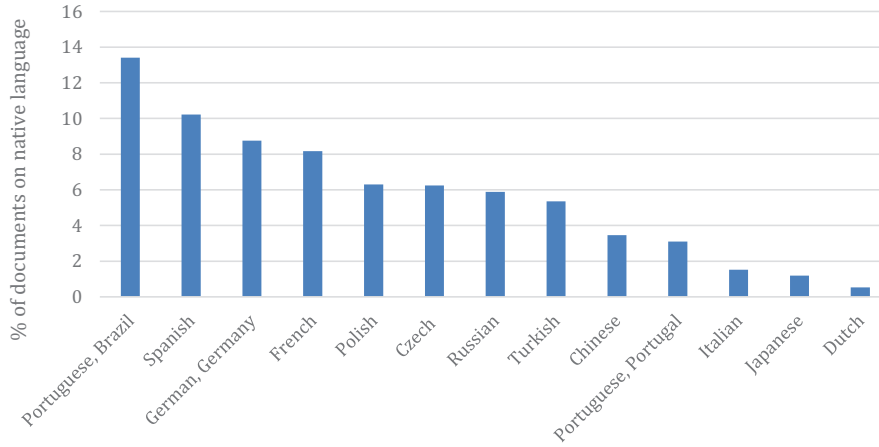


Fig.9 The proportion of documents in native languages for different countries (Web of Science CC, 2006-2017)

Native-language publications are cited less than English-language almost in all examined cases (Fig.10), but in Humanities this influence is lower than for STM. In Brazil, for example, the CNCI of Portuguese publications in Humanities is the same or for some years even higher than for total country documents in this area. CNCI of Spanish-language documents in Humanities is 1,76, with 1,54 for all country publications in Humanities. Influence of language on citation rate is described in literature, though the investigations are not numerous (Diekhoff et al., 2013; Fung, 2008; Shu & Larivière, 2015). Together with data on large proportion of non-English language publications in SSCI and AHCI (Fig.3) it can demonstrate the great role of native-language publications in Social Sciences and Humanities.

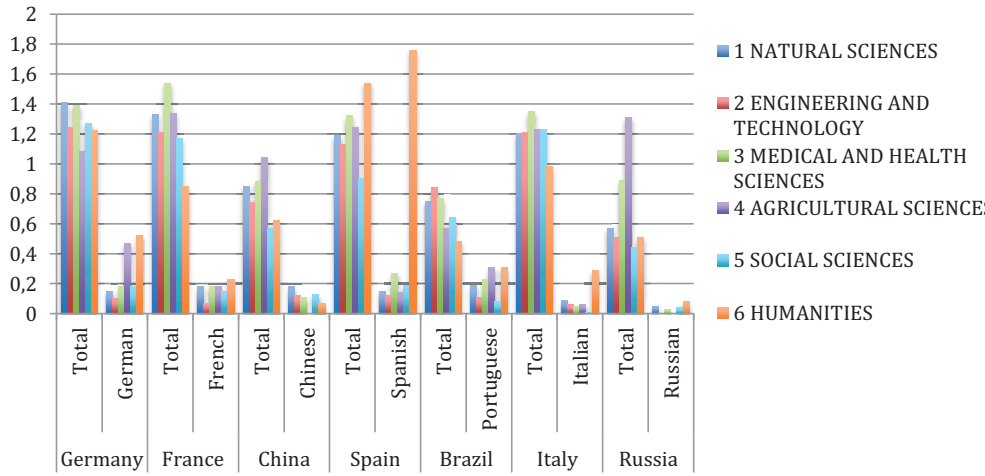


Fig.10 Comparison of CNCI of publications in total and native-language publications (Web of Science, SCI-E, SSCI, AHCI, 2008-2016) for several countries

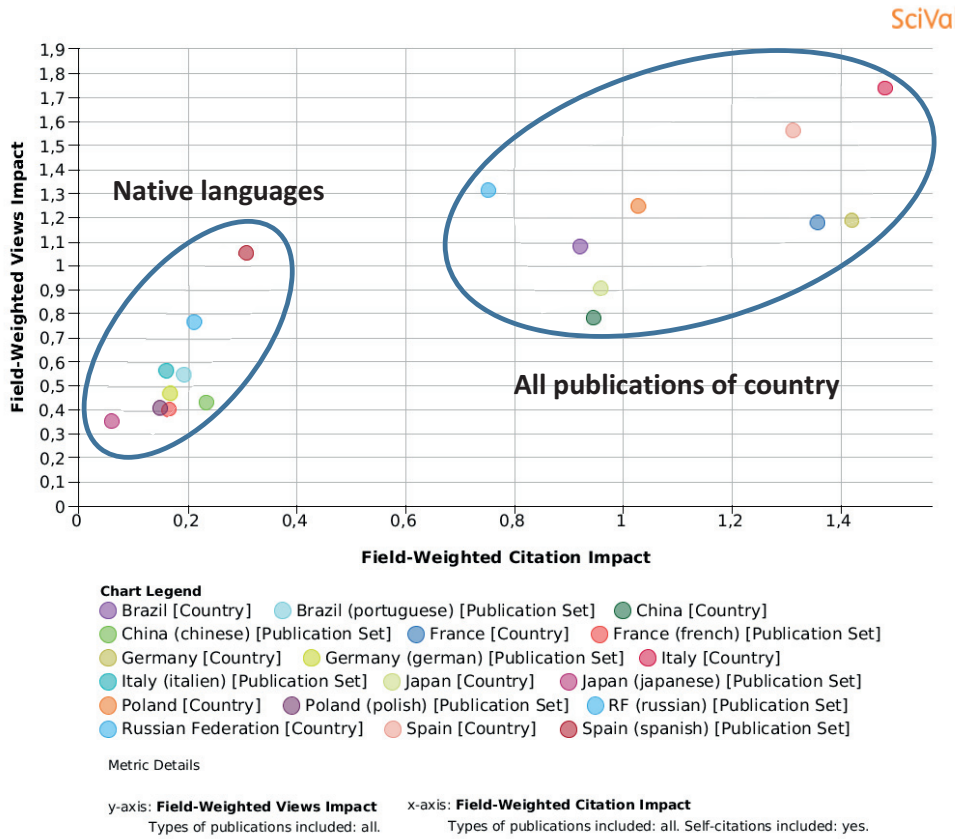


Fig.11 Analysis of Field-Weighted Views Impact and Field-Weighted Citation Impact of countries' publications in total and publications in native language (Scopus, 2015-2016, SciVal data)

The comparison of publications in native languages with all publications of certain countries in Scopus as analyzed in SciVal has demonstrated that Field-Weighted Views Impact and Field-Weighted Citation Impact are significantly lower for native-language publications (Fig.11). In each group we can see a correlation between Field-Weighted Views Impact and Field-Weighted Citation Impact, so we can propose that visibility of publications positively affect their citation. Citations are presumably from the countries where this language is the native one, i.e. citations of Russian-language publications of Russian Federation come from Russia itself and former Soviet Union countries – Belarus, Ukraine, Kazakhstan, etc. It is impossible to analyze the views by countries in SciVal, but we can suppose that their source is similar to that of citations.

Ranking of Multilanguage journals

Several publications demonstrate that English-language journals in total are ranked higher than non-English ones (Liang et al., 2013; Vinther & Rosenberg, 2012). For medical journals the dependence of IF and self-citation level on the number of English-language articles in Multilanguage journals (Diekhoff et al., 2013).

We tried to determine whether the journal ranking (journal IF) correlates with the share of English-language documents in journal for different research areas. We analyzed Multilanguage journals indexed in SCIE and SSCI of Web of Science CC. For the first step we have selected journals marked as Multilanguage from Master Journal List and chose those of them where the number of non-English language documents in 2006-2016 was more than 1. We have found 726 journals that met this criterion. The share of English-language documents in these journals is shown on Fig.12.

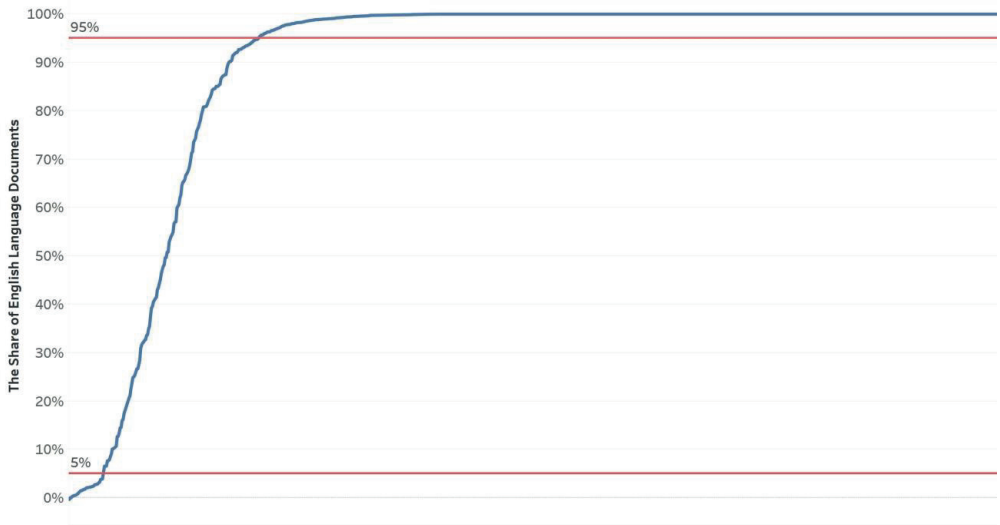
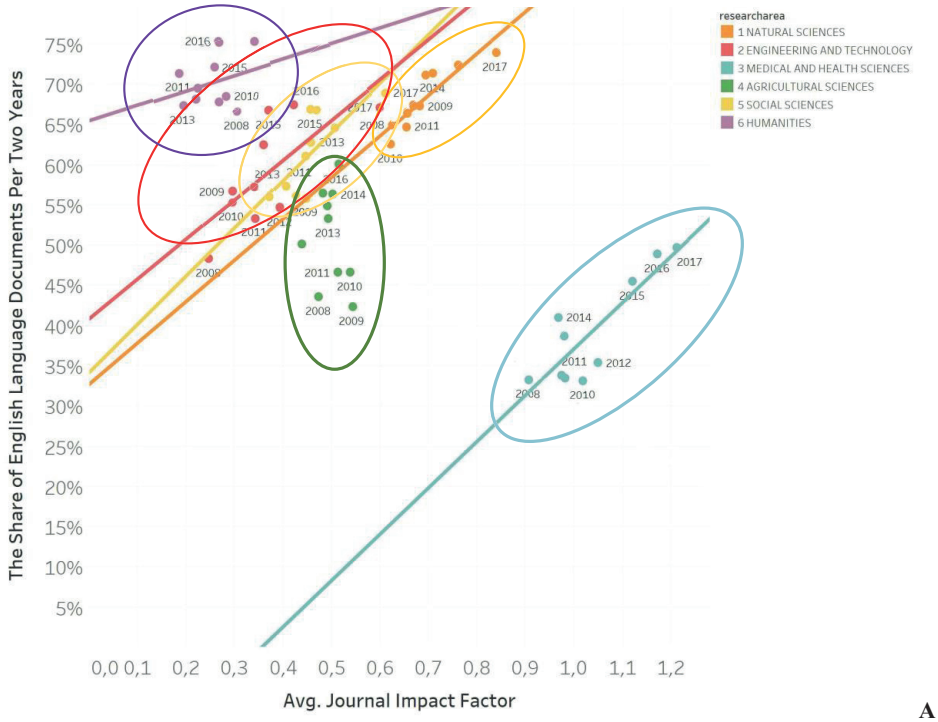


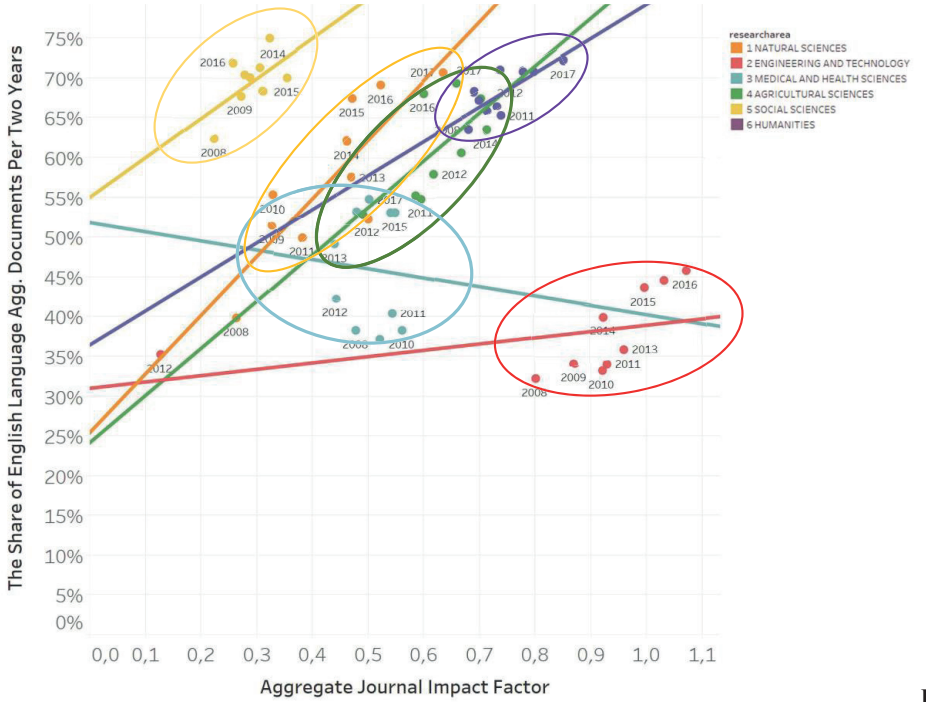
Fig.12 Distribution of Multilanguage journals by proportion of English-language documents (Web of Science CC, 2006-2016).

Later, 159 journals were selected from this list with the share of English-language documents from 5% to 95%. Annual number of English-language documents in 2006-2016 was calculated for each journal to compare these data with journal IF (IF-2017 is to be compared with share of English-language documents in 2015-2016). Selected journals belong to OECD research fields as demonstrated in Table 1.

We found that IF correlates positively with the share of English-language documents in journal not only in medical journals in (IF 2010 and share of English-language documents in 2008-2009) as it had been demonstrated in article mentions above (Diekhoff et al., 2013), but also for other fields of science. Fluctuations of the share of English-language documents during 10-year period influenced the values of IF of journals in natural sciences, engineering & technology and social sciences (Fig.13). Low correlation in humanities and absence of it in agriculture can be explained by peculiarities of citation in these areas and relatively local character of such journals. In agriculture, for example, almost a half of analyzed journals originates from Spanish and Portugal-speaking countries and about 60% of citations came from the same countries.



A



B

Fig.13 Average (A) and aggregate (B) values of journal IF and share of English-language documents by OECD fields of science in 2006-2016.

We understand that such analysis of IF and language dependence is rather crude, so we demonstrate comparison of average and aggregated impact factors for certain journal groups. The certain values for average and aggregate IF for analyzed journals are presented in tables 1(a, b)

Table 1a Number of analyzed journals, average IF coefficients of correlation (Pearson) and P-values for OECD fields of science

OECD field of science	Journal number	R^2	R	p-value
1 Natural Sciences	57	0.816	0,903	0,0003
2 Engineering and Technology	10	0.577	0,760	0,0200
3 Medical Sciences	31	0.773	0,879	0,0016
4 Agricultural Sciences	22	0.042	0,205	0,5692
5 Social Sciences	31	0.577	0,760	0,0108
6 Humanities	8	0.089	0,298	0,4037

Table 1b Number of analyzed journals, aggregate IF coefficients of correlation (Pearson) and P-values for OECD fields of science

OECD field of science	Journal number	R2	R	p-value
1 Natural Sciences	57	0.6045	0.778	0.0081
2 Engineering and Technology	10	0.6937	0,833	0.0027
3 Medical Sciences	31	0.1754	0,419	0.2282
4 Agricultural Sciences	22	0.0048	0,0693	0.8489
5 Social Sciences	31	0.5447	0,738	0.01480
6 Humanities	8	0.3097	0,556	0,095

Full list of analyzed journals, their distribution by OECD fields of science and annual details on IF and English-language share are available on Mendeley Data (Akoev, 2018).

Conclusion

Our results suggest the following:

- The role of English language in scientific publications increases evidently
- Native languages play essential role in social sciences and humanities
- Native-language publications are read and cited less than English-language outside the origin country
- The role of OA (Gold OA or Green OA) for non-English language publications (both views and citation) is even higher than average.
- The journal ranking for Multilanguage journals correlates with the share of English-language publications.

Acknowledgments.

These results for Russian publications were presented on the annual conference of Scientific Electronic Library eLIBRARU.RU “Science Online – XXI” in Austria (27/01/2018-03/02/2018) and were partially published in Russian in journal “Universitetskaya Kniga” (Moskaleva & Akoev, 2018).

References

Akoev, M. (2018). The list of Multi-Language Journals from Web of Science Core Collection masterlist with Impact Factor and Share of English Language Documents In Previous Two Years. Mendeley Data, v.2, DOI: 10.17632/r6dcf26mvp.2

Ammon, U. (2012). Linguistic inequality and its effects on participation in scientific discourse and on global knowledge accumulation – With a closer look at the problems of the second-rank language communities. *Applied Linguistics Review*, *3*(2), 333–355. DOI: 10.1515/applirev-2012-0016

Clarivate Analytics Partners with the Egyptian Knowledge Bank to Power the First Arabic Citation Index - Clarivate. (2018). Retrieved May 30, 2018, from https://clarivate.com/blog/news/clarivate-analytics-partners-egyptian-knowledge-bank-power-first-arabic-citation-index/?utm_source=linkedin&utm_medium=social&utm_campaign=EKB_ABM_ARCI_and_Professional_Services_SAR_EM_Egypt_2018

Di Bitetti, M. S., & Ferreras, J. A. (2017). Publish (in English) or perish: The effect on citation rate of using languages other than English in scientific publications. *Ambio*. DOI: 10.1007/s13280-016-0820-7

Diekhoff, T., Schlattmann, P., & Dewey, M. (2013). Impact of Article Language in Multi-Language Medical Journals - a Bibliometric Analysis of Self-Citations and Impact Factor. *PLoS ONE*, *8*(10), e76816. DOI: 10.1371/journal.pone.0076816

Ethnologue. (2015). Summary by language size: Language Size. *SIL International*. Retrieved from https://www.ethnologue.com/statistics/size

Fung, I. C. H. (2008). Citation of non-English peer review publications - Some Chinese examples. *Emerging Themes in Epidemiology*. DOI: 10.1186/1742-7622-5-12

González-Alcaide, G., Valderrama-Zurián, J. C., & Aleixandre-Benavent, R. (2012). The Impact Factor in non-English-speaking countries. *Scientometrics*, *92*(2), 297–311. DOI: 10.1007/s11192-012-0692-y

Gordin, M. D. (2015). *Scientific Babel. How Science Was Done Before and After Global English. The University of Chicago Press*.

Kulczycki, E., Engels, T. C. E., Pölönen, J., Bruun, K., Dušková, M., et al. (2018). Publication patterns in the social sciences and humanities: evidence from eight European countries. *Scientometrics*, *116*(1), 463-486. DOI: 10.1007/s11192-018-2711-0

Kumar, N., Panwar, Y., & Mahesh, G. (2016). A snapshot of research papers in non - English languages. *CURRENT SCIENCE*, *111*(1), 9–10. Retrieved from https://en.wikipedia.org/w/index.php?title=

Liang, L., Rousseau, R., & Zhong, Z. (2013). Non-English journals and papers in physics and chemistry: bias in citations? *Scientometrics*, *95*(1), 333–350. DOI: 10.1007/s11192-012-0828-0

Liu, F., Hu, G., Tang, L., & Liu, W. (2018). The penalty of containing more non-English articles. *Scientometrics*, *114*(1), 359–366. DOI: 10.1007/s11192-017-2577-6

Liu, W. (2017). The changing role of non-English papers in scholarly communication: Evidence from Web of Science's three journal citation indexes. *Learned Publishing*, *30*(2), 115–123. DOI: 10.1002/leap.1089

Meneghini, R., & Packer, A. L. (2007). Is there science beyond English? *EMBO Reports*, *8*(2), 112–116.

Montgomery, S. L. (2016). Impacts of a Global Language on Science: Are There Disadvantages? In *Language as a Scientific Tool. Shaping Scientific Language Across Time and National Traditions.* (pp. 199–218). Routledge. DOI: 10.4324/9781315657257

Moskaleva, O., & Akoev, M. (2018). Publications on different languages in citation indexes, or If there is a chance for Russian language in science? *Universitetskaya Kniga*, (3), 42–45.

Moskaleva, O., Pislyakov, V., Sterligov, I., Akoev, M., & Shabanova, S. (2018). Russian index of Science citation: Overview and review. *Scientometrics*. DOI: 10.1007/s11192-018-2758-y

Pislyakov, V. (2007). What Are National Citation Indices Needed For ? *Scientific and Technical Libraries*, *2*, 66–68.

Shu, F., & Larivière, V. (2015). Chinese-language articles are biased in citations. *Journal of Informetrics*, *9*(3), 526–528. DOI: 10.1016/j.joi.2015.05.005

Sīle, L., Guns, R., Sivertsen, G., & Engels, T. C. E. (2017). European databases and repositories for Social Sciences and Humanities research output, (July). DOI: 10.6084/m9.figshare.5172322

UNESCO. (2016). *Unesco Science Report*. Retrieved from http://unesdoc.unesco.org/images/0023/002354/235406e.pdf

Vinther, S., & Rosenberg, J. (2012). Impact factor trends for general medical journals: non-English-language journals are lacking behind. *Swiss Medical Weekly 2012 142:3940*, *142*(3940). DOI: 10.4414/smw.2012.13572

Elaborations on a cluster analytical approach to author bibliographic coupling analysis in the context of science mapping

Bo Jarneving[1]

[1] *bo.Jarneving@ub.gu.se*
University of Gothenburg, Gothenburg University Library, Box 222
SE 405 30 Gothenburg (Sweden)

Abstract

Previous research on bibliographic coupling has focused on the research article as the analysed unit. In later research the author has been put forward as the analysed unit and the term author bibliographic coupling analysis (ABCA) has been coined. In the context of science mapping, factor analysis and network analysis have been applied in combination with various methods of counting author bibliographic coupling units. These method approaches are contrasted with a cluster analytical one, applying two recognized similarity measures, the cosine similarity measure and the Jaccard index, in combination with two different cluster methods, the SLM-algorithm and Ward's method. The purpose of this approach was to test the suitability of ABCA as a tool for deriving classifications in the context of science mapping. Findings revealed that the cosine similarity measure provided with the most valid cluster solution for both cluster methods. The optimal method combination comprised Ward's method and the cosine similarity measure. However, the evaluation of clusters' quality disclosed a lack of structure in data, suggesting that cluster analysis is not a suitable method in the context of ABCA. Considering the features of author bibliographic coupling data, the suitability of ABCA in the context of science mapping was challenged.

Introduction

In the evolution of science mapping, most of the significant fields in bibliographic records have been used for the purpose of mapping one or another aspect of the scientific enterprise. Mainly bibliometric models based on the co-occurrence principle have been applied, using keywords, cited references, as well as authors' names and institutional addresses. A deviation from this pattern is the analysis of shared references between analysed units – bibliographic coupling analysis. Most of the research on bibliographic coupling has focused on the mapping of subject fields and research fronts applying the research article as the analysed unit. Later research has applied the authors of the research articles as the unit of analysis and the term author bibliographic coupling analysis (ABCA) has been coined (Zhao & Strotmann, 2008b). Here, various ways of counting common author-references have been applied in combination with factor analysis and network analysis (Zhao & Strotmann, 2008b; Ma, 2012). The objective of the present study was to test the applicability of ABCA for science mapping when applied to cluster analytical methods.

The remainder of this paper is organized as follows: continuing under *Introduction*, previous research is presented. Under *Data and methods*, data sources and methods of data selection are accounted for, as well as the cluster analysis and the applied proximity measures. *Research design and rationale* concludes the data and methods section. Under *Findings* the empirical results are presented and a summary of the same is given under *Summary of findings*. The implications of findings and methods are finally elaborated on under *Discussion*.

Previous research

At the end of the '50s and in the early '60s Kessler introduced the concept of bibliographic coupling through a series of MIT-reports and journal articles (Kessler, 1958, 1960, 1962, 1963a, 1963b). In an early MIT-report from Lincoln Laboratories, Kessler (1962) described bibliographic coupling: "[a] single item of reference shared by two documents is defined as a unit of coupling between them". Kessler developed this brief description further in terms of two graded criteria:

Criterion A – A number of articles constitute a related group G_A if each member of the group has at least one reference (one coupling unit) in common with a given test article P_o. The coupling strength between P_o and any member of G_A is measured by the number of coupling units between them. G^n_A is that portion of G_A that is linked to P_o through n coupling units.

Criterion B – A number of articles constitute a related group G_B if each member of the group has at least one coupling unit with every other member of the group. The coupling strength of G_B is measured by the number of coupling units between its members.

Kessler clearly associated bibliographic coupling with the core issues of information science: classification, indexing and information retrieval. Kessler found that meaningful partitions of article groups could be accomplished on a small scale (1962) and subsequently successfully scaled up the research setting to several thousand papers (1963 a, b). Kessler also compared bibliographic coupling with subject indexing (1965) and found a strong correlation between the methods. However, all these experiments were confined to the indexing-retrieval context and to the field of physics, why the applicability of the method in a more complex (multidisciplinary) research setting was yet to be tested. In 1984 Vladutz and Cook performed a large-scale test of the relation between bibliographic coupling and subject relatedness, using the 1981 annual Science Citation Index. They concluded that results confirmed that valid results with respect to the subject relatedness was achieved, and that using bibliographic coupling on a large citation database was practically feasible. A year before, Sen and Gan (1983) presented a purely theoretical paper on bibliographic coupling connecting with Kessler's idea of G_B groups but here referred to as 'cliques'. Importantly, they suggested an improvement of the bibliographic coupling strength, presenting the coupling angle (CA), which is the cosine of the angle for two vectors of cited references.

Subsequent research covered, amongst other things, subject relatedness focusing on G_A groups and term-relatedness (Peters, Braam and van Raan, 1995), the identification of hot topics through so called 'core documents', i.e., current, strongly coupled and well-integrated papers in the research front (Glänzel and Czerwon, 1996; Jarneving, 2007; Glänzel and Thijs, 2012) and comparison of co-citation with bibliographic coupling (Persson, 1994; Jarneving, 2005).

Analogous to the branching of the co-citation method into author co-citation analysis (ACA), a paper suggesting the concept of author bibliographic coupling analysis (ABCA) was presented by Zhao and Strotman (2008b). Elaborating on how to measure ABCA, they suggested that an author's references should be weighted by the frequency of occurrence, acknowledging the importance of repeated referencing of papers. Their method was tested applying factor analysis for the mapping of the field of information science over two five-year periods, and in a sequel over a later five-year period (Zhao & Strotman, 2014). Another empirical study applying ABCA for science mapping purposes was presented by Ma (2012). Here, a mapping of Chinese Library and Information Science was pursued applying several different methods for the counting of shared references. Research applying ABCA in the context of grant peer-review was presented by Sandström and Wang in 2015. In this paper the notion of 'cognitive distance' was defined in terms of both text data and author bibliographic coupling. Finally, the relation between author bibliographic coupling strength and citation exchange was explored by Gazni and Didegah (2016) over eighteen subject areas.

Data and methods

The selection of data was guided by the ambition to facilitate the interpretation of results and to connect to previous research. Therefore, the 'classic' journal set used by White & McCain in their seminal paper from 1998 on ACA and mapping of information science was applied.

This journal set has been applied and referred to in subsequent papers on science mapping (e.g. Persson, 2001; Zhao & Strotmann 2008a, 2008b). A total of 4186 papers, articles and reviews, distributed over 12 selected journals were downloaded from Web of Science (Table 1).

Table 1. The distribution of papers over selected journals.

# papers	*Journal*
902	Scientometrics
537	Journal of the American Society for Information Science and Technology
497	Information Processing & Management
437	Journal of Information Science
410	Electronic Library
278	Journal of Documentation
256	Information Technology and Libraries
201	Library Resources & Technical Services
188	Library & Information Science Research
365	ASIST
91	Annual Review of Information Science and Technology
24	Program-Automated Library and Information Systems

In total 113,820 cited references were counted and the average number of references per paper was 27. The number of unique references was 76,790. Inspired by Colliander & Ahlgren (2012), an effort was made to unify similar references using normalized Levenshtein distances (Levenshtein, 1966). This implied that reference pairs with a normalized Levenshtein distance ≥ 0.95 were unified to one version and that the number of unique references decreased to 76,095.

With the aim of finding well integrated authors, the following selection method was applied. All author pairs with only a single bibliographic coupling unit were filtered out and the median was computed on basis of the remaining pairs. Applying the median as a threshold, author pairs with a coupling strength at or below the median were filtered out. On basis if this new distribution, the number of bibliographic coupling links for each author was computed. Finally, from a list of the 100 most productive authors, those 50 authors with most bibliographic coupling links at the applied threshold of coupling strength were selected.

Proximity measures

In this study, two proximity measures were applied: the cosine similarity measure and the Jaccard index. A proximity measure reflects the similarity or dissimilarity between two objects, hence the notion of similarities and dissimilarities. In this sense, one may consider counts of bibliographic coupling units as similarities (cf. Leydesdorff, 2006). In the bibliometric literature, different ways of counting author bibliographic couplings have been put forward. In Rousseau (2010) the simple author bibliographic coupling strength was presented and in Zhao and Strotmann (2008b) the so called 'minimum' method. Also, Ma (2012) empirically tested an additional third method labelled the 'combined' method. The following definitions are based on those papers:

- Minimum method: for a paper *x* cited by both A and B, its frequency of citation by A and B respectively is computed, and the lowest frequency is selected as its weight. The sum of such weights is then the bibliographic coupling strength for AB.
- Combined method: for a paper *x* cited by both A and B, its frequency of citation by A and B respectively is computed, and the product of these frequencies is its weight. The sum of such weights is then the bibliographic coupling strength for AB.

- Simple method: for two authors, A and B, the number of unique papers cited by both is their bibliographic coupling strength

Notably, the combined method allows for pairs of factors of (much) varying size giving rise to the same final weight and the simple method disregards re-citation.

Though proximities sometimes can be obtained directly from observations of phenomena, they are commonly derived from a raw-data matrix containing the data necessary for computing the proximities. Conceptually, the proximities are derived from an asymmetric $m \cdot n$ data matrix and contained in a symmetric $n \cdot n$ proximity matrix. In the case of ABCA, the data matrix is a reference-author matrix with m rows and n columns and the resulting proximity matrix an $n \cdot n$ author-author matrix. Sometimes, proximity matrices are transformed to yet another, but final proximity matrix based on how pairs of objects relate to other objects in the population under study (cf. McCain, 1990). This approach is sometimes referred to as a 'global' approach (Ahlgren, Jarneving & Rousseau, 2003).

The statistical literature presents us with a plethora of proximity measures, but only a few have been applied on a more common basis in citation-based science mapping. The Pearson product-moment correlation has often been the preferred proximity measure in ACA (White & Griffith, 1981; McCain, 1990, White & McCain, 1998). As accounted for by McCain (1990), the Pearson product moment correlation functions as a measure of how similar author profiles are and removes differences in scale between authors. The use of this measure in ACA has been criticized on grounds of not being invariant to the adding of zeros to vectors of author co-citation counts and the cosine measure (amongst others) suggested as an alternative (Ahlgren et al., (2003).

The cosine similarity measure is often referred to as the Salton's cosine formula in the literature because of its association with information retrieval research (cf. Baeza-Yates & Ribero-Neto, 1999; Salton, 1983). The cosine similarity measure has been applied repeatedly in bibliographic coupling analyses (e.g. Glänzel & Czerwon, 1996; Jarneving, 2007) and in ABCA quite recently (Sandström & Wang, 2015). It computes the cosine of the angle between two vectors x and y and is defined as

$$sim(x, y) = \frac{x \cdot y}{\|x\| \|y\|}$$

where $\|x\|$ is the Euclidean norm of vector x, $\|y\|$ the Euclidean norm of vector y and $x \cdot y$ the scalar product of x and y. In Leydesdorff (2008), analysing author co-citation data, the author compared the cosine similarity measure with the Jaccard index and concluded that the cosine measure was superior when applied on an asymmetric data matrix. Notably, both the Jaccard index and the cosine similarity measure are invariant to adding zeros, hence sparse matrices would not cause problems. The Jaccard index is a binary similarity measure which is defined for two objects, A and B, as

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|}$$

where the nominator is the size of the intersection and the denominator the size of the union. Notably, the Jaccard coefficient is related to the simple method of counting while the cosine measure, as applied, involve weighted citations (the frequency of occurrence).

The cluster analysis

Two different cluster methods were applied: Ward's method (Ward, 1963) and the smart local moving (SLM) algorithm (Waltman, L., & Van Eck, 2013). In addition, two evaluative methods were applied: one for the comparison of partitions and one for the cluster quality.

Ward's method is based on the size of an error sum of squares criterion where the objective at each stage is to minimize the increase of the total within cluster error sum of squares (Everitt, Landau & Leese, 2001, pp.57). Ward's method requires dissimilarity coefficients, why a transformation from similarities to dissimilarities was needed. As similarity is the complement to dissimilarity when measured in the range [0,1], values on the cosine similarity measure and the Jaccard index were converted by subtracting them from 1. In order to find the optimal number of clusters, Wards method was applied in combination with Mojena's 'Upper Tail Rule' (Mojena, 1977). The 'Upper Tail Rule' evaluates the distribution of clustering criterion values at each level of fusion and tries to find an unusually large gap in an increasing series of such values (Wishart, 2005).

The SLM algorithm applies a local moving heuristic for the optimization of modularity (i.e. structure and denseness of a network). The basic idea of this heuristic is to repeatedly move nodes from one community (cluster) to another in a way that increases the modularity using a quality function (Waltman, L., & Van Eck, 2013; Waltman, Boyack, Colavizza and van Eck, 2017). In addition, a resolution parameter that affects the granularity level (i.e. number of clusters) was used. For the evaluation of cluster quality in terms of internal cohesiveness and external separation, the Silhouette index (s) was applied. This method works as follows: for each object i there is a corresponding value $a(i)$ which denotes the average dissimilarity of i to all other objects in cluster a to which i belongs. Next, consider any other cluster of the partition and the average dissimilarity of i to all objects in that cluster and denote this $d(i,C)$. Finally, select the $d(i,C)$ with the smallest dissimilarity with i and denote this $b(i)$. Then $s(i)$ is arrived at combining $a(i)$ and $b(i)$ as follows:

$$s(i) = \begin{cases} 1 - a(i)/b(i) \text{ if } a(i) < b(i) \\ 0 \text{ if } a(i) = b(i), \\ b(i)/a(i) - 1 \text{ if } a(i) > b(i) \end{cases}$$

Also, when i is the only member in a, $s(i)$ is set to zero (Rousseeuw, 1987).

In order to evaluate the agreement between partitions accomplished by different methods, the adjusted Rand index was applied. This index takes on values between 0 and 1 and when two partitions agree completely it reaches its maximum value. Its precursor, the Rand index, is based on the agreement or disagreement of every pair of n objects. The index computes the proportion of the total number of pairs that agree, that is, are in the same cluster according to the first partition and in the same cluster according to the second partition *or* in different clusters according to the first partition and in different cluster according to the second partition. However, this index tends to increase by the number of clusters, hence the need for an index corrected for chance – the adjusted Rand index (Everitt et al., 2001, pp.181).

Finally, applying a heuristic method for the labelling of clusters, titles, keywords and abstracts from authors' articles published during the period of observation were examined. In addition to personal knowledge, this facilitated the assignment of one subfield to each author. With the intention of identifying topics in subfields, the title of each author's most cited paper during the period of observation was added to the cluster display.

The research design and rationale

In order to investigate the problem from different directions, two cluster methods were applied for the evaluation of two proximity measures, the overall objective being to explore the

appropriateness of ABCA as a science mapping method. There exist quite many plausible cluster methods while considerations of general features of bibliometric data imply a more delimited set of appropriate choices of proximity measures.

Both the cosine similarity measure and the Jaccard index have been frequently applied in previous science mapping enterprises. Both proximities take on values in the interval [0,1] and the maximum value is reached when two vectors are identical. Otherwise they are quite different as the Jaccard index is 'the intersection over union' while the cosine measure is the 'cosine of the angle of two vectors'. As previously noted, both measures are invariant to adding zeros, which means that they may be properly applied on a sparse reference-author matrix. This, however, is not the case with the Pearson product moment correlation which accordingly was ruled out as a plausible proximity measure despite its frequent use in bibliometric studies. The issue of sparse matrices also implies the exclusion of standard proximity measures of the Euclidean type as the co-absence of a reference bear no useful meaning. Should one apply counts of shared references, using any of the aforementioned counting methods, this would entail a severe loss of information as only references in the intersection of a pair of reference-author vectors would be used. Hence, this type of similarity measure was also considered non-optimal in the current context.

The hitherto commonly used cluster methods in science mapping pertain to the group of hierarchical agglomerative methods. From the author's own empirical experience, Ward's method usually performs better than several other hierarchical agglomerative methods. One defining feature of hierarchical methods is that once two clusters are merged, this cannot be undone. Such methods may be contrasted with optimization methods where the number of random starts and the number of iterations may be varied. One such is the SLM-algorithm presented by Waltman and Van Eck (2013).

Findings

It could be seen that the choice of cluster algorithm influenced the partitioning considerably. Measuring the agreement between partitions when the cosine similarity measure was applied, the value on the adjusted Rand index was 0.44 and 0.38 when the Jaccard index was applied. Applying the SLM-algorithm and setting the resolution parameter to 1.0 (for standard modularity-based community detection), the highest modularity value was attained but also the smallest overall Silhouette widths as well as the largest clusters, merging different subfields. The largest overall Silhouette widths were found with the resolution parameter set to 3.0. At this level, the SLM-method produced somewhat more fragmented partitions than Ward's method in terms of average cluster size (Table 2). Applying the cosine similarity measure, clusters with a larger overall Silhouette width showed up for both cluster methods. The largest overall Silhouette width is seen for the combination of Ward's method and the cosine similarity measure.

Table 2. Evaluation of clusters generated by two cluster methods and two similarity measures.

Cluster method	*Similarity measure*	*Overall Silhouette width*	*# of clusters*	*Average cluster size*
SLM	Cosine similarity	0.19	15	3.33
	Jaccard index	0.12	17	2.94
Ward	Cosine similarity	0.24	13	3.85
	Jaccard index	0.13	12	4.17

It is quite clear that co-authorships influenced the result. This is most notable in terms of author relations resisting changes of methods; hence some authors co-occur in the same cluster in all

partitions. Examples of such author relations are 'Bassecoulard-Zitt' and 'Egghe-Rousseau'. In a similar sense, a couple of clusters emerged in all partitions: {Harries G, Tang R, Thelwall M, Vaughan L, Wilkinson D}, {Ellis D, Ford N, Foster A, Spink A, Wilson TD}. It could be concluded that all method combinations by and large distinguished between subfields such as bibliometrics, information retrieval, information seeking and webometrics. Considering the intellectual overlap between some subfields and the fact that authors may publish in more than one subfield, a few transgressions from this general pattern is not surprising.

As mentioned, the combination of Ward's method with the cosine similarity measure resulted in the largest overall Silhouette width. Still, authors in cluster 4, 7 and 8 have been assigned very small Silhouette widths, meaning that they in a sense are misclassified (Table 3). Following Kaufman and Rousseeuw (2005, p.88) an overall Silhouette width ≤ 0.25 would imply that no substantial structure has been found. Hence, the overall Silhouette width of 0.24 for this cluster solution would suggest that an artefact has been produced. However, the suggested threshold is based on rule by thumb and perhaps not equally relevant in all research settings. But even if we accepted this somewhat lower overall Silhouette width (0.24), 44 percent of the authors would still be considered misclassified. Allowing the intellectual content of an author's most cited paper published during the period of observation guide the assessment of cluster topics, the variation of research themes within clusters suggests that those clusters with a larger Silhouette width make more sense than those with a lower width. For instance, the composition of the larger Cluster 8 makes little sense while the theme of Cluster 1 is evident. It seems that the Silhouette width is largely in agreement with a subjective assessment.

Table 3. The cosine similarity measure applied to Ward's method of clustering: Silhouette width, subfield and the most cited paper.

Cluster	*Silhouette*	*Author*	*Subfield*	*Most cited paper during the period of observation*
1	0.41	Ellis D	Information seeking	Modelling the information seeking patterns of engineers and research scientists in…
1	0.36	Ford N	Information seeking	Serendipity and information seeking: an empirical study
1	0.45	Foster A	Information seeking	Serendipity and information seeking: an empirical study
1	0.27	Spink A	Information seeking	Real life, real users, and real needs: a study and analysis of user queries on the web
1	0.44	Wilson TD	Information seeking	Models in information behaviour research
2	0.12	Bilal D	Information seeking	Children's use of the Yahooligans! Web search engine: I. Cognitive, physical, and…
2	0.18	Wang PL	Information seeking	Users' interaction with World Wide Web resources: an exploratory study using…
3	0.53	Beheshti J	Information seeking	Gender differences in collaborative Web searching behaviour: an elementary school study
3	0.28	Cole C	Information seeking	Information and poverty: information-seeking channels used by African American …
3	0.47	Large A	Information seeking	Gender differences in collaborative Web searching behaviour: an elementary school study
4	-0.09	Beaulieu M	Information retrieval	Experimentation as a way of life: Okapi at TREC
4	0.01	Ingwersen P	Information retrieval	Cognitive perspectives of information retrieval interaction: Elements of a cognitive IR theory
4	0.03	Jarvelin K	Information retrieval	Collaborative Information Retrieval in an information-intensive domain
4	-0.03	Toms EG	Information seeking	Usability of the academic library Web site: Implications for design
4	0.00	Vakkari P	Information seeking	Task-based information searching
5	0.29	Chowdhury GG	Bibliometrics	Bibliometric cartography of information retrieval research by using co-word analysis
5	0.26	Foo S	Bibliometrics	Bibliometric cartography of information retrieval research by using co-word analysis
6	0.26	Wolfram D	Webometrics	Searching the Web: The public and their queries
6	0.27	Zhang J	Information retrieval	Internet search engines' response to metadata Dublin Core
7	-0.03	Aoe J	Information retrieval	Documents similarity measurement using field association terms
7	0.01	Chen HC	Information retrieval	Internet browsing and searching: User evaluations of category map and concept space techniques
7	-0.02	Choi KS	Information retrieval	A comparison of collocation-based similarity measures in query expansion
7	0.04	Mostafa J	Information retrieval	Automatic classification using supervised learning in a medical document filtering application
7	-0.01	Savoy J	Information retrieval	A stemming procedure and stopword list for general French corpora
8	-0.29	Bar-Ilan J	Webometrics	Data collection methods on the Web for informetric purposes - A review and analysis
8	-0.04	Chen CM	Bibliometrics	Visualizing knowledge domains
8	-0.02	Cronin B	Bibliometrics	Hyperauthorship: A postmodern perversion or evidence of a structural shift in scholarly co…
8	-0.06	Furner J	Bibliometrics	Scholarly communication and bibliometrics

Table 3. (continued).

8	*-0.16*	Glänzel W	Bibliometrics	National characteristics in international scientific co-authorship relations
8	-0.09	Gupta BM	Bibliometrics	A comparison of productivity of male and female scientists of CSIR
8	-0.02	Hjorland B	LIS-theory	Domain analysis in information science - Eleven approaches - traditional as well as innovative
8	-0.07	Leydesdorff L	Bibliometrics	Theories of citation?
8	-0.03	Oppenheim C	Bibliometrics	The correlation between citation counts and the 1992 research assessment exercise ratings…
8	-0.02	White HD	Bibliometrics	Visualizing a discipline: An author co-citation analysis of information science, 1972-1995
8	-0.10	Wilson CS	Bibliometrics	The literature of bibliometrics, scientometrics, and informetrics
8	-0.14	Vinkler P	Bibliometrics	Relations of relative scientometric impact indicators. The relative publication strategy index
9	0.41	Egghe L	Bibliometrics	Methods for accrediting publications to authors or countries: Consequences for evaluation studies
9	0.40	Rousseau R	Bibliometrics	Social network analysis: a powerful strategy, also for the information sciences
10	0.56	Robertson SE	Information retrieval	Experimentation as a way of life: Okapi at TREC
10	0.57	Spärck-Jones K	Information retrieval	A probabilistic model of information retrieval: development and comparative experiments…
11	0.47	Moed HF	Bibliometrics	Journal impact measures in bibliometric research
11	0.52	van Leeuwen TN	Bibliometrics	Language biases in the coverage of the Science Citation Index and its consequences for …
11	0.35	Van Raan AFJ	Bibliometrics	Fatal attraction: Conceptual and methodological problems in the ranking of universities by …
12	0.97	Bassecoulard E	Bibliometrics	Shadows of the past in international cooperation: Collaboration profiles of the top five producers…
12	0.97	Zitt M	Bibliometrics	Shadows of the past in international cooperation: Collaboration profiles of the top five producers…
13	0.66	Harries G	Webometrics	Motivations for academic web site interlinking: evidence for the Web as a novel source of…
13	0.54	Tang R	Information retrieval	Toward an understanding of the dynamics of relevance judgment: An analysis of one person's search…
13	0.75	Thelwall M	Webometrics	Search engine coverage bias: evidence and possible causes
13	0.52	Vaughan L	Webometrics	Search engine coverage bias: evidence and possible causes
13	0.65	Wilkinson D	Webometrics	Motivations for academic web site interlinking: evidence for the Web as a novel source…

Summary of findings

The empirical test showed that the cosine similarity measure produced better partitions than the Jaccard index over two different cluster algorithms. The agreement between cluster solutions produced by the SLM algorithm respectively Ward's method was low and the SLM algorithm generated more fragmented cluster solutions. The optimal combination of cluster method and similarity measure was the combination of Ward's method and the cosine similarity measure. The optimal cluster solution provided with an interpretable sectioning of authors over subfields though the composition of clusters with a low Silhouette width generally made less sense. In total, 44 % of the authors were regarded misclassified. Conclusively, the applied methods failed to generate an adequate mapping of the field under study.

Discussion

In the present study findings showed that no grouping of data could be justified. Comparing the cluster analysis with a corresponding factor analysis[1] it was seen that only 35 out of 50 authors were assigned a factor loading of at least 0.4, which again points to a lack of structure in data. Hence, the main issue to elaborate further is to what extent author bibliographic coupling data generally provide with enough structure for science mapping purposes. When author names serve as proxies for the cognitive content of sets of publications, as in ACA and ABCA, things get complicated. As noted by Zhao and Strotman (2008b), well established authors may expand their research interests and publish on several topics but are mainly cited with regard to one or two topics. This, the authors suggest, may explain why co-citation counts are more concentrated than author bibliographic coupling counts. In their study this phenomenon showed up as "significantly better model fits" when factor analysing author co-citation counts. Hence, when references are amalgamated to the author level, more topic spread could be expected, as opposed to the document level where references reflect more delineated areas of interest. In this sense, authors are not ideal representatives of research areas. ABCA is probably a proper method in the context of research evaluation and science communication but its usefulness as a method for science mapping could be challenged. Also, further testing of the method should involve fields other than information science.

Concerning the issue of co-authorship. Rousseau (2010) recommend that co-authored papers should be excluded when computing the bibliographic coupling strength between co-authors. As demonstrated in this study, co-authorship may have a strong influence on results and breaking up strong co-author relations would generate a different mapping. One may, however, question if this would be in line with a general purpose of mapping the cognitive associations between authors.

References

Ahlgren, P., Jarneving, B. & Rousseau, R. (2003). Requirements for a co-citation Similarity Measure, with Special Reference to Pearson's Correlation Coefficient. *Journal of the America Society for Information Science*, 54(6):550–560.

Baeza-Yates, R. A., & Ribeiro-Neto, B. A. (1999). Modern information retrieval. Harlow: Addison-Wesley. Modelling (Chapter 2).

[1] The extraction method was principal component analysis and the rotation method oblimin with Kaiser normalization. Factor loadings were derived from the pattern matrix in SPSS.

Colliander, C., Ahlgren, P. (2012). Experimental comparison of first and second-order similarities in a scientometric context. *Scientometrics*, 90(2), 675–685.

Everitt, B., Landau, S., & Leese, M. (2001). *Cluster analysis* (4th ed.). London: Arnold. Measurement of proximity (Chapter 3).

Gazni, A. & Didegah, F. (2016). The relationship between authors' bibliographic coupling and citation exchange: analyzing disciplinary differences. *Scientometrics*, 107(2):609–626.

Glänzel, W. & Czerwon, H.J. (1996). A new methodological approach to bibliographic coupling and its application to the national, regional and institutional level. *Scientometrics*, 37(2):195-221.

Glänzel, W. & Thijs, B. (2012). Using 'core documents' for detecting and labelling new emerging topics. *Scientometrics* (2012) 91:399–416.

Jarneving, B. (2005). A comparison of two bibliometric methods for mapping of the research front. *Scientometrics*, 65(2): 245-263.

Jarneving, B. (2007). Bibliographic coupling and its application to research-front and other core documents. *Journal of Informetrics*, 1(4):287-307.

Kaufman, L. & Rousseeuw, P.J. (2005). *Finding groups in data*. New Jersey: Wiley-Interscience. Introduction (Chapter 1).

Kessler M.M. (1958). Concerning some problems of intrascience communication. Massachusetts Institute for Technology, Lincoln laboratory.

Kessler M.M. (1960). An experimental communication center for scientific and technical information. Massachusetts Institute for Technology, Lincoln laboratory.

Kessler M.M. (1962). An experimental study of bibliographic coupling between technical papers. Massachusetts Institute for Technology, Lincoln laboratory.

Kessler M.M. (1963a). Bibliographic coupling between scientific papers. *American Documentation*, 14(1):10-25.

Kessler M.M. (1963b). Bibliographic coupling extended in time: Ten case histories. *Information Storage and retrieval*, 1(4):169-187.

Kessler M.M. (1965). Comparisons of the results of bibliographic coupling and analytical subject indexing. *American Documentation*, 16(3):223-233.

Levenshtein, VI. (1966). Binary codes capable of correcting deletions, insertions, and reversals. *Dok Akad Nauk SSSR*, 163:845-848.

Leydesdorff, L. (2006). Co-occurrence matrices and their applications in information science: extending ACA to the web environment. *Journal of the American Society for Information Science and Technology*, 57(12):1616-1628.

Leydesdorff, L. (2008). On the Normalization and Visualization of Author Co-Citation Data: Salton's Cosine versus the Jaccard Index. *Journal of the American Society for Information Science and Technology*, 59(1):77–85.

Ma, R. (2012). Author bibliographic coupling analysis: A test based on a Chinese academic database. *Journal of Informetrics*. 6(4):532-542.

McCain K.W. (1990). Mapping authors in intellectual space: A technical overview. *Journal of the American Society for Information Science and Technology*, 41(6):433-443.

Mojena, R. (1977). Hierarchical grouping methods and stopping rules: An evaluation. *Computer Journal*, 20(4):359–363.

Persson, O. (1994). The intellectual base and research fronts of JASIS 1986-1990. *Journal of the American Society for Information Science and Technology*, 45(1):31-38.

Persson, O. (2001). All author citations versus first author citations. *Scientometrics*, 50(2):339-344.

Peters, H.P.F., Braam, R. R. & van Raan, A.F.J. (1995). Cognitive resemblance and citation relations in chemical engineering publications. Journal of the American Society for Information Science, 46(1):9-21.

Rousseeuw, P.J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational an Applied Mathematics*, 20(November):53-65.

Rousseau, R. (2010). Bibliographic coupling and co-citation as dual notions. In B. Larssen (Ed.), A festschrift in honour of Peter Ingwersen, special volume of the e-zine of the international society for scientometrics and informetrics, June 2010.

Salton, G., & McGill, M. J. (1983). Introduction to modern information retrieval. New York: McGraw-Hill.

Sandström, U. & Wang, Q. (2015). Defining the role of cognitive distance in the peer review process with an explorative study of a grant scheme in infection biology. *Research Evaluation* 24 (3):271–281.

Sen, S.K. & Gan, S.K. (1983). A mathematical extension of the idea of bibliographic coupling and its applications. *Annals of Library Science and Documentation*, 30(2):78-82.

Vladutz, G. & Cook, J. (1984). Bibliographic coupling and subject relatedness. Proceedings of the ASIS Annual Meeting, 47:204-207.

Waltman, L., & Van Eck, N.J. (2013). A smart local moving algorithm for large-scale modularity-based community detection. *European Physical Journal B*, 86(11): 471.

Waltman L., Boyack K.W., Colavizza G. & Van Eck N.J. (2017). A principled methodology for comparing relatedness measures for clustering publications. 16th International Conference of the International Society for Scientometrics and Informetrics, Wuhan. In: Proceedings of the 16th International Conference of the International Society for Scientometrics and Informetrics. 691-702.

Ward, J.H. (1963). Hierarchical Grouping to Optimize an Objective Function. *Journal of the American Statistical Association*, 58(301):236-244.

White, H.D. & Griffith, B. (1981). Author co-citation: a literature measure of intellectual structure. *Journal of the American Society for Information Science*, 32(3):163-171.

White, H.D., McCain, K.W. (1998). Visualizing a discipline: An author co-citation analysis of information science, 1972-1995. *Journal of the American Society for Information Science*, 49(4):327-335.

Wishart, D. (2005). Number of clusters. In B. Everitt, & D. D. Howell (Eds.), Encyclopedia of statistics in behavioral science (pp. 1442–1446). Chichester: John Wiley & Sons.

Zhao, D., & Strotmann, A. (2008a). Information science in the first decade of the Web: An enriched author co-citation analysis. *Journal of the American Society for Information Science and Technology*, 59(6):916–937.

Zhao, D., & Strotmann, A. (2008b). Evolution of Research Activities and Intellectual Influences in Information Science 1996–2005: Introducing Author Bibliographic-Coupling Analysis. *Journal of the American Society for Information Science and Technology*, 59(13), 2070–2086.

Zhao, D., & Strotmann, A. (2014).The Knowledge Base and Research Front of Information Science 2006–2010: An Author co-citation and Bibliographic Coupling Analysis. *Journal of the American Society for Information Science and Technology*, 65(5):995–1006.

The anatomy of retracted papers in the Web of Science, 1998-2017

Philip Roe[1] and Grant Lewison [1,2]

[1] *philip@evaluametrics.co.uk, grantlewison@aol.co.uk*
Evaluametrics Ltd, 157 Verulam Road, St Albans, AL3 4DW (UK)

2grant.lewison@kcl.ac.uk
King's College London, Guy's Hospital, Great Maze Pond, London SE1 9RT (UK)

Abstract

This paper (work-in-progress) provides preliminary results from an analysis of papers with the word "retracted " in their titles in a 20-year period. There were 5566 such papers, averaging 0.02% of all Web of Science documents, but with an anomalous peak in 2011 caused by the retraction of a set of conference proceedings from Singapore. More than half the other retracted papers were in clinical medicine or biomedical research, so their original publication may have caused harm to patients. The countries with the highest percentages of retractions (other than in these conference proceedings) were Iran, Pakistan, Egypt, Algeria and Saudi Arabia, all Muslim-majority countries. Future work will examine the possible reasons for retraction in these countries.

Introduction

While researching the outputs of papers on cancer from Middle East and North African countries, we noticed that a number of papers from Iran had titles that began with the word RETRACTED in upper case letters. This was intriguing, and we wondered how widespread the practice of retraction was in the last two decades. There have been several high-profile cases where authors have failed to disclose the funding they had received from pharmaceutical companies, or fabricated results, several of which have been reported in leading newspapers (Sample, 2017; Srivastava, 2018), and recently the Swedish government sacked the entire board of the prestigious Karolinska Institutet because of their failure to deal with an Italian surgeon (Hawkes, 2016) who had faked his results. A search on the Web of Science (WoS, © Clarivate Analytics) soon showed that there were literally thousands of retracted papers. A search for papers with the word "retracted" in their titles identified them quite easily, but there were a few papers in two other categories: papers that reviewed the subject, usually from a limited perspective, and papers that dealt with retraction as a subject in its own right, such as the retraction of aircraft doors (Fu *et al*, 2016), or of cancer-associated fibroblasts (Stadler *et al.*, 2017).

This paper surveys the field in order to show how widespread the practice of retraction is, which countries are primarily responsible, in which subjects retractions occur relatively most frequently and to identify some leading retractors. The prior literature mainly concerns medical papers (Rada, 2005; Samp *et al.,* 2012; Peterson, 2013; Inoue & Muto, 2016; Rai & Sabharwal, 2017). Brainard and You (2018) surveyed the situation recently, and drew attention to the database *Retraction Watch*, created by Ivan Oransky and Adam Marcus, which covers over 18,000 retracted papers and provides reasons for the retraction of most of them. About half were seen as fraudulent in some way (because of fabrication, falsification or plagiarism), but many were the result of honest mistakes, or failure to reveal sources of financial support. Retractions may be initiated by the journals concerned, or by sharp-eyed journalists, and this greater vigilance seems to have reduced the time interval between publication and retraction (Steen *et al*., 2013).

For the present analysis, it was of course necessary to normalise the results with respect to all papers in the 20-year period chosen for study. Our intention is not to find fault so much as to point out geographical and subject areas where more attention to the reproducibility of the

results may be needed to ensure that the scientific enterprise continues to command widespread confidence and support. It should be emphasised at the outset that retraction remains rather rare, with fewer than 0.02% of papers normally being retracted, except for the year 2011, when the number was unusually high, see Figure 1, below.

Methodology

Bibliographic data on all articles and reviews covered in the WoS (all six databases) during the 20 calendar years 1998-2017 with the word *retracted* in their titles were downloaded to a series of text files, and converted to an MS Excel spreadsheet by means of a visual basic application (VBA) program (macro) written by PR. The file of 5857 papers contained all the papers' addresses, which were parsed by another macro to show the fractional counts of each country. Inspection of the titles and document types revealed which papers were not actually retractions, but fell in one of the two other categories mentioned in the introduction. There were 10 papers reviewing the field of retracted papers, and 276 (4.7%) where the actual subject matter of the paper was retraction.

During the same years, the WoS covered a total of 23,964,509 papers and these, as well as the retracted set, were analysed by means of the built-in database software to show their publication year, their allocation to subject areas (based on the journal in which they were published), and the countries of their authors (whole or integer counts). The percentages of retracted papers in each of these three categories were then calculated so as to normalise the analyses of the retracted papers. We also attempted to identify the leading retractors, although this was sometimes very difficult, particularly with Chinese names where there were many homonyms.

Results

The numbers of all WoS papers that were retracted, year by year, are shown in Figure 1. There is a spike in the number for 2011, which appears to have reached a high of 0.08% of the total, but in other years it averaged 0.02%, with a gradual rise to 2011, and then a decline to 0.01% in 2017. The spike in 2011 was caused by the withdrawal of all 762 of the papers in the printed volume of the 2011 International Conference on Energy and Environmental Science (ICEES 2011), *Energy Procedia*, Volume 11. The conference was held in Singapore, and was "removed by the publisher due to insufficient assurances by the programme organisers that the professional ethical codes of publishing and standards were applied consistently" (https://www.sciencedirect.com/science/article/pii/S1876610214004536). Despite this mass retraction, the 762 conference papers still managed to attract many citations (140), as did the 31 flawed papers by Hendrik Schön (Luwel et al., 2018), see Table 3, below.

Retracted paper subject areas

The WoS lists several hundred subject areas, but there is considerable overlap between them. The top two subject areas, Energy & Fuels and Environmental Sciences, with very large numbers of retracted papers, were affected because of the withdrawal of the papers of the 2011 International Conference on Energy and Environmental Science, discussed above, which is classified under both these subject areas. Apart from the top two subject areas, many of the others are biomedical in character, and the retracted papers could have caused damage to patients who were diagnosed wrongly or given inappropriate treatment.

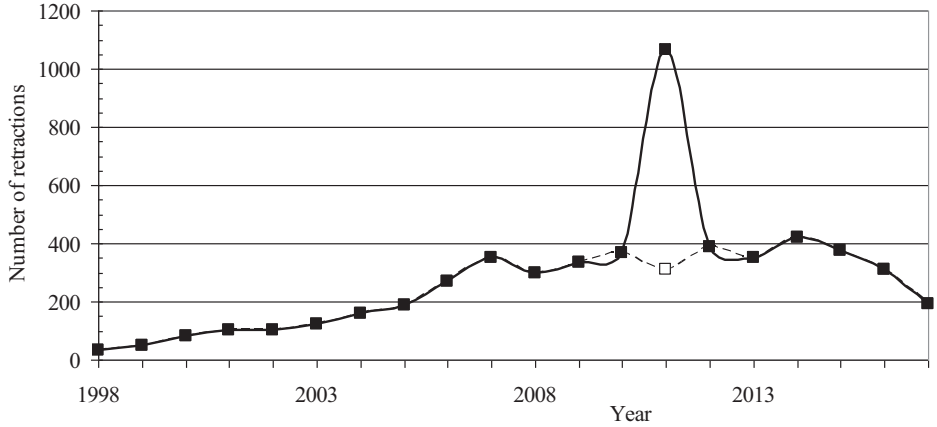


Figure 1. Number of WoS papers that were subsequently retracted, 1998-2017. The open square shows the number without the retracted conference proceedings on Energy and Environmental Science in 2011.

Table 1.WoS subject areas with at least 100 retractions. Biomedical subject areas in bold.

Web of Science Categories	*N*	*%*	*Web of Science Categories*	*N*	*%*
Energy & Fuels	839	0.16	Mathematics Applied	107	0.02
Environmental Sciences	852	0.12	**Immunology**	142	0.02
Oncology	496	0.05	**Pharmacology Pharmacy**	184	0.02
Medicine Research Experimental	210	0.04	**Biotechnology Applied Microbiology**	108	0.02
Cell Biology	319	0.04	**Surgery**	184	0.02
Multidisciplinary Sciences	256	0.04	**Endocrinology Metabolism**	105	0.02
Biochemistry Molecular Biology	495	0.04	**Neurosciences**	176	0.02
Mathematics	156	0.03	**Medicine General Internal**	107	0.01

The countries with the most retractions

These are ranked by the percentage of their papers that have been retracted, see Table 2. It is noticeable that the countries at the head of the list, in the first column, include nine out of 13 listed that have a Muslim majority and are in the Organisation of Islamic Cooperation, and seven in East Asia out of the nine that are listed here (some are both). However, the data for China have been inflated because of the retraction of the 2011 International Conference on Energy and Environmental Science, for which the country contributed almost every single paper (n = 755). Aside from this mass retraction, Chinese researchers only retracted 1223 papers, or 0.031% of their total output.

The leading authors who have retracted papers

These are listed in Table 3, with their institutional addresses and research subject areas. (Half of these authors were also listed by Brainard and You (2018).) Several of the authors are clearly colleagues in the same institution, and it may be that the errors or behaviour that led to the papers being retracted were not their individual fault. What the table does show is that

there are clearly some authors who have had to retract rather large numbers of papers. In total, there were 17,610 different names, although some of them may have referred to the same individual who used different sets of initials on different papers. For example, Abbruzzese-J with three papers is clearly the same as Abbruzzese-JL with two.

Table 2. Countries with at least 30,000 papers in the WoS in 1998-2017, ranked by percentage of papers retracted. Members of the Organisation of Islamic Cooperation (OIC) shown in bold, those in East Asia shown in *italics*.

Countries	*N*	*%*	*Countries*	*N*	*%*	*Countries*	*N*	*%*
Iran	204	0.054	USA	1307	0.012	Denmark	23	0.007
China	1978	0.051	**Morocco**	5	0.012	Brazil	49	0.007
Pakistan	47	0.046	**U Arab Emirates**	4	0.012	Czech Rep.	17	0.007
Egypt	60	0.041	Israel	35	0.011	Hungary	10	0.007
Algeria	18	0.039	*Vietnam*	4	0.011	South Africa	13	0.006
Saudi Arabia	41	0.031	Romania	20	0.011	Austria	20	0.006
Tunisia	19	0.029	Sweden	58	0.011	France	102	0.006
India	305	0.028	**Nigeria**	5	0.010	Slovakia	5	0.006
Serbia	19	0.025	Germany	247	0.010	Argentina	9	0.005
South Korea	199	0.021	Italy	144	0.010	Poland	24	0.005
Singapore	48	0.020	Greece	25	0.009	Colombia	3	0.004
Malaysia	35	0.019	Norway	22	0.009	Belgium	19	0.004
Japan	390	0.018	Spain	104	0.009	Slovenia	3	0.004
Taiwan	81	0.015	Australia	99	0.009	New Zealand	7	0.004
Ireland	26	0.015	Switzerland	53	0.009	Mexico	9	0.004
Turkey	74	0.015	United Kingdom	239	0.009	Chile	4	0.003
Croatia	11	0.015	Finland	23	0.009	Bulgaria	2	0.003
Netherlands	116	0.014	***Indonesia***	5	0.008	Estonia	1	0.003
Thailand	18	0.014	Canada	111	0.008	Russia	21	0.003
Bangladesh	4	0.013	Portugal	19	0.008	Ukraine	2	0.002

Discussion

Retraction of scientific papers is a relatively rare event, averaging one in approximately 5000 papers, but it can cause major upsets in the careers of the scientists involved. It is particularly egregious for medical and biomedical papers, which are often the source of new knowledge for clinicians and underpin the recommendations in clinical guidelines. Based on the classification of papers into the CHI major fields (Narin *et al*., 1976), there were 2025 papers in Clinical Medicine and a further 1012 in Biomedical Research, together over half the total, which may have had an incorrect impact on human (or animal) health.

Table 3. List of the authors who appear to have retracted the most papers in 1998-2017, with the number of their retracted papers, their most recent addresses, years of publication, and subject area.

Authors	*N*	*Addresses*	*Years*	*Subject*
Fujii, Yoshitaka	64	Univ Tsukuba, Inst Clin Med, Dept Anesthesiol, Tsukuba, Ibaraki 305, Japan	1998-2008	Anaesthesia
Zhong, Hua	42	Jinggangshan Univ, Coll Chem & Chem Engn, Prov Key Lab Coordinat Chem, Jian 343009, China	2006-07	Crystallography
Stapel, Diederik A	39	Tilburg Univ, Tiber, Social Psychol Dept, NL 5000 LE Tilburg, Netherlands	2006-11	Mood and social feelings
Boldt, Joachim	37	Klinikum Stadt Ludwigshafen, Dept Anesthesiol & Intens Care Med, D-67063 Ludwigshafen, Germany	1998-2011	Anaesthesia, hydroxyethyl starch
Sarkar, Fazlul H.	34	Wayne State Univ, Dept Pathol, Karmanos Canc Inst, Sch Med, Detroit, MI 48201, USA	2006-13	Cancer cells
Yang, Xuemei	34	Jinggangshan Univ, Coll Chem & Chem Engn, Prov Key Lab Coordinat Chem, Jian 343009, China	2006-07	Crystallography
Tanaka-H	33	Univ Tsukuba, Inst Clin Med, Dept Anesthesiol, Tsukuba, Ibaraki 3058576, Japan	1998-2005	Post-operative nausea
Wang, Zhiwei	30	Wayne State Univ, Sch Med, Barbara Ann Karmanos Canc Inst, Dept Pathol, Detroit, MI 48201, USA	2006-13	Cancer cells
Schon-JH	29	Bell Labs, Lucent Technol, Murray Hill, NJ 07974, USA	1998-2002	Transistors and superconductivity
Banerjee, Sanjeev	28	Wayne State Univ, Sch Med, Barbara Ann Karmanos Canc Inst, Dept Pathol, Detroit, MI 48201, USA	2006-13	Cancer cells

The result for 2011 is highly anomalous, and provides a warning against the placement of undue trust in bibliometrics unless the reasons for any unusual results are fully explored.

Table 2 shows that, of the countries publishing at least 1500 papers per year in the study period, the ones with the highest percentages tend to be either Muslim-majority, or East Asian, or both. This result is robust, and does suggest that there is a lower ethical standard for research integrity there compared with, say, North America and Europe. If we base the norm for retractions on the totals for these two continents (0.0117% and 0.0089% respectively, with a mean value of 0.0102%), then all of the countries in the first column of Table 2 are retracting more papers than this average percentage. Tests of statistical significance on the Poisson distribution with one degree of freedom shows that $p << 1\%$ for all of them, except for Thailand, Croatia and Bangladesh, where the numbers are too small to be statistically significant.

Why the number of retractions is relatively so high in Muslim-majority countries, and in east Asia, will need additional investigation. It may be because of the motivation of the researchers, or greater diligence by journalists and other investigators.

References

Brainard, J., You, J. What a massive database of retracted papers reveals about science publishing's 'death penalty'. (2018) *Science*, DOI: 10.1126/science.aav8384

Fu, Y.L., Liu, Y. , & Teng, Y.T. (2016) Electromechanical Actuator System for **Retractable** Door in Aircraft Environment Control System. *3rd International Conference on Mechanical, Industrial, and Manufacturing Engineering (MIME 2016)*, Los Angeles, CA, USA; 145-149.

Hawkes, N. (2016) Swedish government sacks entire Karolinska Institute board over Macchiarini. *BMJ-British Medical Journal,* 354, i4894 DOI: 10.1136/bmj.i4894

Inoue, Y. & Muto, K. (2016) Noncompliance with Human Subjects' Protection Requirements as a Reason for Retracting Papers: Survey of Retraction Notices on Medical Papers Published from 1981 to 2011. *Accountability in Research-Policies and Quality Assurance*, 23(2), 123-135

Luwel, M., van Eck, N.J., van Leelulwen, T. (2018) The Schön case: Analyzing in-text citations to papers before and after retraction. *Proceedings of the STI Conference*, CWTS, Leiden University, The Netherlands: 1025-1030..

Narin, F., Pinski, G & Gee, H.H. (1976) Structure of the Biomedical Literature. *Journal of the Association for Information Science and Technology,* 27(1), 25-45

Peterson, G.M. (2013) Characteristics of Retracted Open Access Biomedical Literature: A Bibliographic Analysis. *Journal of the American Society for Information Science and Technology*, 64(12), 2428-2436

Rada, R. (2005) A case study of a retracted systematic review on interactive health communication applications: Impact on media, scientists, and patients. *Journal of Medical Internet Research*, 7(2)

Rai, R. & Sabharwal, S. (2017) Retracted Publications in Orthopaedics Prevalence, Characteristics, and Trends. *Journal of Bone and Joint Surgery-American Volume*, 99(9)

Samp, J.C., Schumock, G.T. & Pickard, A.S. (2012) Retracted Publications in the Drug Literature. *Pharmacotherapy*, 32(7), 586-595

Sample, I. (2017) Top UK geneticist faces new inquiry over claims of research misconduct. *The Guardian,* 3 January.

Srivastava, R. (2018) One top oncologist failed to disclose payments from big pharma. That hurts all cancer patients. *The Guardian,* 26 September.

Stadler, S., Nguyen, C.H., Schachner, H., Milovanovic, D., Holzner, S., Brenner, S. et al. (2017) Colon cancer cell-derived 12(S)-HETE induces the **retraction** of cancer-associated fibroblast via MLC2, RHO/ROCK and Ca2+ signalling. *Cellular and Molecular Life Sciences*, 74(10), 1907-1921

Steen, R.G., Casadevali, A., Fang, F.C. (2013) Why has the number of scientific retractions increased? *PLoS ONE*, 8(7), e68397.

New Wine in Old Bottles? Examining Institutional Hierarchy in Mobility Networks of Prestigious Awards Laureates, 1901–2017

Fan Jiang[1] and Nian Cai Liu[2]

[1] *fanjiang@sjtu.edu.cn*
Shanghai Jiao Tong University, Graduate School of Education, 800 Dongchuan Road, Shanghai (China)

[2] *ncliu@sjtu.edu.cn*
Shanghai Jiao Tong University, Graduate School of Education, 800 Dongchuan Road, Shanghai (China)

Abstract

One of the major trends in academia today is the constant growth of the mobility of scientists. However, the unbalanced mobility of scientific elites can not only lead to the institutional hierarchy in research sectors but also result in an "elite circulation" phenomenon. In this study, we collect the institutional mobility information of 1351 laureates of 21 prestigious international academic awards and construct the laureate mobility networks. Furthermore, we examine the structure of these networks by using several statistical approaches. We find that there are hierarchical structures in laureate mobility networks, both in different research fields and at different historical periods, indicating that only a few institutions link much more laureates and most institutions at a limited level. In addition, the Gini coefficients showing that a stronger level of institutional inequality in laureate mobility networks over time.

Introduction

The mobility of scientific elites has been acknowledged as a powerful mechanism for knowledge transfer and ideas spread as they bring experience, scientific merit and social ties to their new workplace. Geographically, the mobility of researchers across different countries and regions have become a major policy objective since the term "brain drain/gain" had been used for the first time in 1963 in a report by the Royal Society of London in reference to the exodus of British scientists to the United States (Adams, 1968; Brandi, 2006). Furthermore, from a historical view, Bernal (1954) proposed a theoretical framing of "the shifting of world science activities center", which is later examined correlating with the international and organizational mobility of scientific elites in different time periods (Yuasa, 1962; Zhao & Jiang, 1985; Zuckerman, 1995). On the other hand, the unbalanced mobility of scientific elites has resulted in the academic inequality and social stratification of research sectors, which made some scholars reconsider the drain/gain approach and start talking about "elite circulation" at institutional level (Pareto, 1968; Meyer, 2001; Cañibano, Otamendi & Solís, 2011). As a result, there is a shared understanding that the mobility of top scientists could influence a country or an institution's scientific performance, which further drives researchers and policymakers to find the exact patterns of the evolving structure of academic mobility, and the predictors of mobility scale and directions.

From the network perspective, extensive studies reported that the status hierarchy exists at institutional level across a range of academic fields, such as physics (Gaston, 1970; Cole & Cole, 1973; Clauset et al. 2015), computer science (Clauset, Arbesman & Larremore, 2015; Morgan et al., 2018), information science (Wiggins & Sawyer, 2012; Zhu & Yan, 2017), sociology (Crane, 1969; Burris, 2004), economics (Coupé, 2003; Mixon, Torgler & Upadhyaya, 2017), communication science (Barnett et al., 2010; Mai, Liu& González-Bailón, 2015), and law (Katz et al., 2011; Arewa, Morriss & Henerson, 2014), and at different stages of academic careers, such as PhD exchange (Burris, 2004), faculty hiring (Clauset, Arbesman & Larremore, 2015; Zhu & Yan, 2017), postdoc employment (Melin, 2005; Zubieta, 2009; Cantwell, 2011) and job transfer (Cruz-Castro & Sanz-Menéndez, 2010; Schlagberger, Bornmann & Bauer, 2016). The core finding of above studies is that there is a stratified network where only a few

institutions producing and attracting most academic elites. However, the institutional hierarchy and inequality from a mobility network's view were rarely discussed.

In this study, we construct the directed, weighted mobility networks of 1351 laureates of 21 prestigious international academic awards from 1901 to 2017, and further investigate the structures of them. The main contributions of our study are twofold. First, we examine network-based measures derived from laureate mobility networks to understand their implications on understanding institutional hierarchy in different research fields and how this hierarchy structure evolving at different historical periods. Second, we examine the level of institutional inequality in laureate mobility networks through Lorenz curve and Gini coefficient.

Methods

Working hypotheses

To better organize our empirical analyses, we make the following working hypotheses: (1) The laureate mobility networks are characterized by hierarchical structures. In other words, there are a few well-linked institutions (hubs) with many laureates while majority of institutions (spokes) only have a limited number of linkages; (2) The laureate mobility networks become more linked meanwhile hub institutions remain better connected over time; and (3) There are high levels of institutional inequality in laureate mobility networks and become stronger over time. As our analysis is exploratory, we will use these working hypotheses as the starting point for our examination.

Data

We collected 1351 laureates of 21 international academic awards between the awards' inception and the awarding year 2017. These 21 awards are the most prestigious ones in each subject from the previous reputation survey (see Zheng & Liu, 2015; Jiang & Liu, 2018). Table 1 provides the information of these 21 awards. Then, from these laureates' CV data and additionally with Wikipedia information, we collect the institutions of where the laureates obtained their highest degree and received the award. Thus, we identify a total number of 437 institutions of these laureates' mobility affiliation.

Table 1. The information of 21 prestigious international academic awards.

Field	Subject	Award name	First awarding year	Number of laureates
Natural sciences	Astronomy	Crafoord Prize in Astronomy	1985	11
	Chemistry	Nobel Prize in Chemistry	1901	178
	Earth sciences	Crafoord Prize in Geosciences	1983	14
	Mathematics	The Abel Prize	2003	18
	Physics	Nobel Prize in Physics	1901	207
	Physiology & Medicine	Nobel Prize in Physiology or Medicine	1901	214
Engineering sciences	Chemical engineering	R. H. Wilhelm Award	1966	47
	Civil engineering	Freyssinet Medal	1970	28
	Electrical & Computer engineering	A. M. Turing Award	1966	67
	Energy engineering	Eni Award	1987	118
	Environmental engineering	Tyler Prize	1974	72
	Material engineering	Von Hippel Award	1976	41
	Mechanical engineering	ASME Medal	1921	85
Social	Business & Management	AOM Distinguished Scholarly	1983	35

sciences		Contributions to Management Award		
	Communication sciences	ICA Steven H. Chaffee Award	2000	17
	Economics	The Sveriges Riksbank Prize in Economic Sciences in Memory of Alfred Nobel	1969	79
	Education	AERA Distinguished Contributions to Research in Education Award	1964	57
	Law	The Stockholm Prize in Criminology	2006	21
	Public administration	APSA Herbert A. Simon Book Award	2005	16
	Political sciences	The Johan Skytte Prize	1995	24
	Sociology	ISA Award for Excellence in Research and Practice	2014	1

Network representation

To these data, we construct the directed, weighted mobility networks based on these laureates from where they received their highest degrees to their award-winning institutions. Each node represents an institution, the size of node representing the centrality of an institution, and the width of the linkage indicating the number and direction of laureates' mobility between two institutions. Figure 1 shows an example of five laureate mobility network.

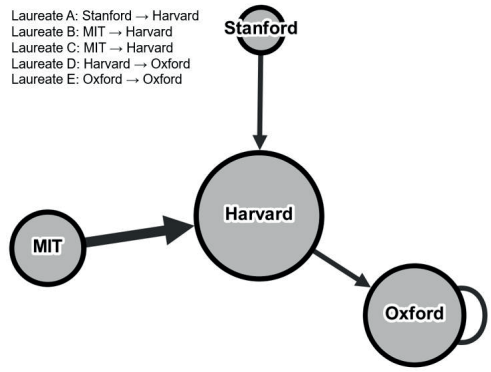


Figure 1. An example of a directed, weighted laureate mobility network.

As such, we construct the mobility networks in different research fields (natural sciences, engineering sciences, and social sciences), and at different time periods (from 1901 to 1940, 1980, and 2017) respectively. In this study, we use Fruchterman-Reingold layout in Gephi software to visualize the networks.

Network structure measures

The starting point for evaluating network structures is to introduce a set of fundamental network metrics, such as degree, degree distribution, and centrality value of the nodes. Degree is the most basic index to measure the statistical property of nodes in a network, while the degree distribution helps to explore the underlying processes by which the network has come into existence. Centrality analysis measures the importance of each node and provides the ranking of nodes in a network. Below we formally specify each of these measures:

(1) Degree and degree distribution. The degree k_i of node i is defined as the number of nodes it is directly connected to, and is given by:

$$k_i = \sum_{j \in V(i)} a_{ij}$$

where $V(i)$ denotes the neighbour set of node i. For a scale-free network which is generated by a preferential attachment rule, the degree distribution $P(k)$ follows a power law, given by:

$$P(k) = k^{-\alpha}$$

where α is the fitted power-law exponent (Clauset, Shalizi & Newman, 2009).

(2) Centrality analysis. In scientometrics, PageRank and its variants have been used to evaluate the importance of different objects such as papers, scientists, journals and institutions (Gleich, 2015; Zhou et al., 2018). In this study, we use PageRank algorithm to identify the centrality of institutions in laureate mobility networks. The standard PageRank algorithm (Brin & Page, 1998) in directed and unweighted networks can be defined as:

$$PR(p_i) = \frac{1-d}{n} + d \sum_{p_j \in \mathcal{M}(p_i)} \frac{PR(p_j)}{D_{out}(p_j)}$$

where $PR(p_i)$ represents the PageRank value of node p_i , n is the total number of nodes in the given network, $\mathcal{M}(p_i)$ denotes the adjacency matrix of node p_i , d (=0.85) is the damping factor. As our networks are directed and weighted, so we justify the algorithm as:

$$PR(p_i) = \frac{1-d}{n} + d \sum_{p_j \in \mathcal{M}(p_i)} \frac{w(p_j) \times PR(p_j)}{D_{out}(p_j)}$$

where $w(p_j)$ is the cumulative degree of node p_j.

Lorenz curve and Gini coefficient

Lorenz curve (Gastwirth, 1971) is a measure that studies the inequality of wealth distribution in economics. It is represented as a curve drawn in a two-dimensional graph, in which the relationship between population and income is plotted. Along with a Lorenz curve, the Gini coefficient (Gastwirth, 1972) is another widely used measure of inequality. The Gini coefficient is a value between zero and one, in which zero represents perfect equality, and one expresses absolute inequality. In this study, we use Lorenz curve to represent as a function between the cumulative fraction of laureates and the cumulative fraction of laureates' mobility institutions. Then, the Lorenz curve, along with the Gini coefficient, explains the level of institutional inequality in different research fields and at different time periods. The definition of the Gini coefficient is as follows:

$$G = \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} |x_i - x_j|}{2n^2 \bar{x}}$$

where G represents the Gini coefficient of the laureate mobility network, x is an observed value, n is the number of values observed, and x *bar* ($\bar{x}$) is the mean value.

From the perspective of Economics, it is commonly regarded Gini coefficient less than 0.2 as absolutely equality or too fair, 0.2~0.3 as comparative equality, 0.3~0.4 as relatively equality, 0.4~0.5 as relatively inequality, above 0.5 as strong inequality (Zhang & Huang 2018).

Results

Statistical properties of laureate mobility networks

As can be seen in Figure 2, the laureate mobility networks' P(k) fit the power-law function with scaling exponents of α = 1.74 in Natural sciences, α = 1.66 in Engineering sciences, and α = 1.47 in Social sciences. Thus, these characterize the laureate mobility networks as scale-free

networks according to Barabási and Albert's (1999) definition. Therefore, a node that is caused by statistical fluctuations will receive more linkages than others during the initial stages and will increasingly get more linkages, becoming a hub. Similarly, poorly connected nodes tend to continue with a low level of linkages. In other words, there is a hierarchy or structural inequality among these networks. In addition, the PageRank value distributions also fit the power-law function with scaling exponents of α = 1.71 in Natural sciences, α = 1.51 in Engineering sciences, and α = 1.59 in Social sciences.

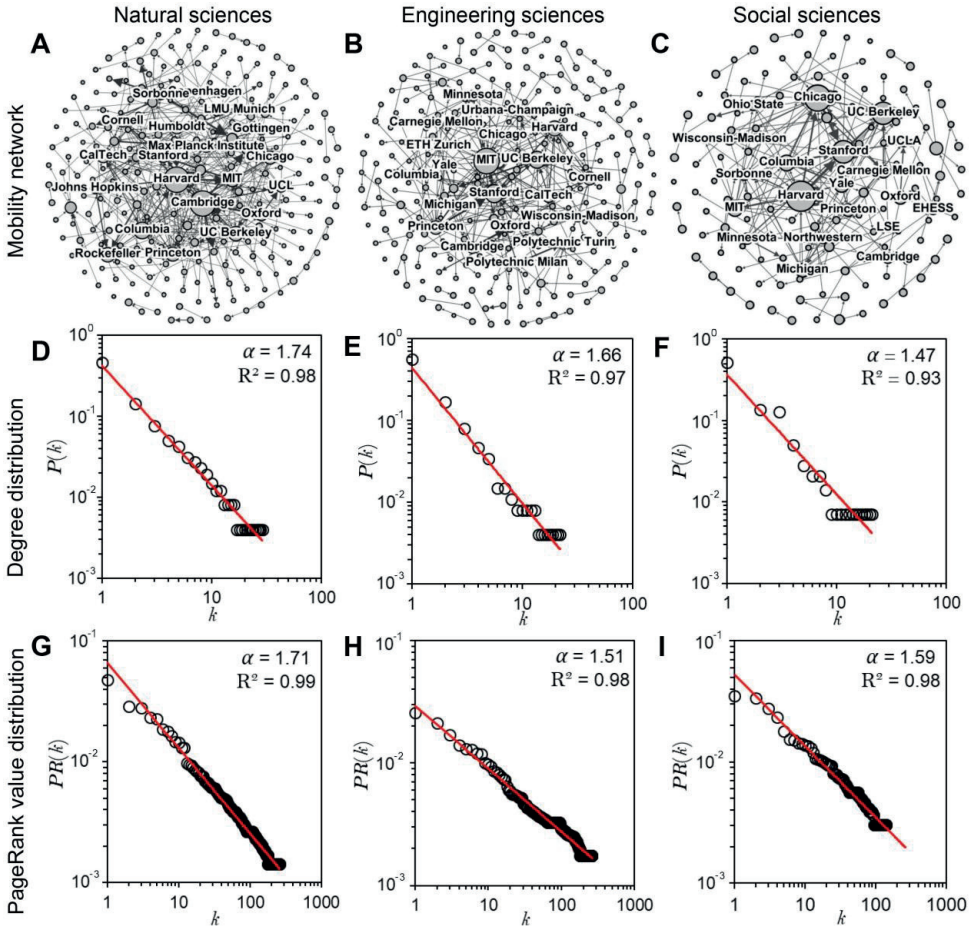


Figure 2. The laureate mobility networks in natural sciences (A), engineering sciences (B), and social sciences (C), their degree distributions (D, E, and F), and their PageRank value distributions (G, H, and I).

Table 2 lists the top five hub institutions in each research field and their statistical properties, $N(k)$ demonstrates the name ID of institutions, $D(k)$ represents the weighted degree of an institution, and $PR(k)$ indicates its PageRank value. In natural sciences, University of Cambridge is the most hub institution, while MIT in engineering sciences, and Harvard University in social sciences.

Table 2. The top five hub institutions in different research fields.

Natural sciences			Engineering sciences			Social sciences		
$N(k)$	$D(k)$	$PR(k)$	$N(k)$	$D(k)$	$PR(k)$	$N(k)$	$D(k)$	$PR(k)$
Cambridge	91	0.047	MIT	58	0.026	Harvard	49	0.035
Harvard	71	0.039	Stanford	45	0.021	Chicago	39	0.033
Columbia	40	0.038	UC Berkeley	52	0.017	Stanford	33	0.028
Rockefeller	22	0.031	Cambridge	31	0.014	UC Berkeley	26	0.023
UC Berkeley	35	0.029	Harvard	28	0.013	Columbia	15	0.018

Hub institutions over time

Another important question is, do hub institutions remain better linked over time, or do hub institutions have been changed? In Figure 3, we can easily find that the total number of linkages of networks keep growing, meanwhile, there are hub institutions which have most of linkages in each network.

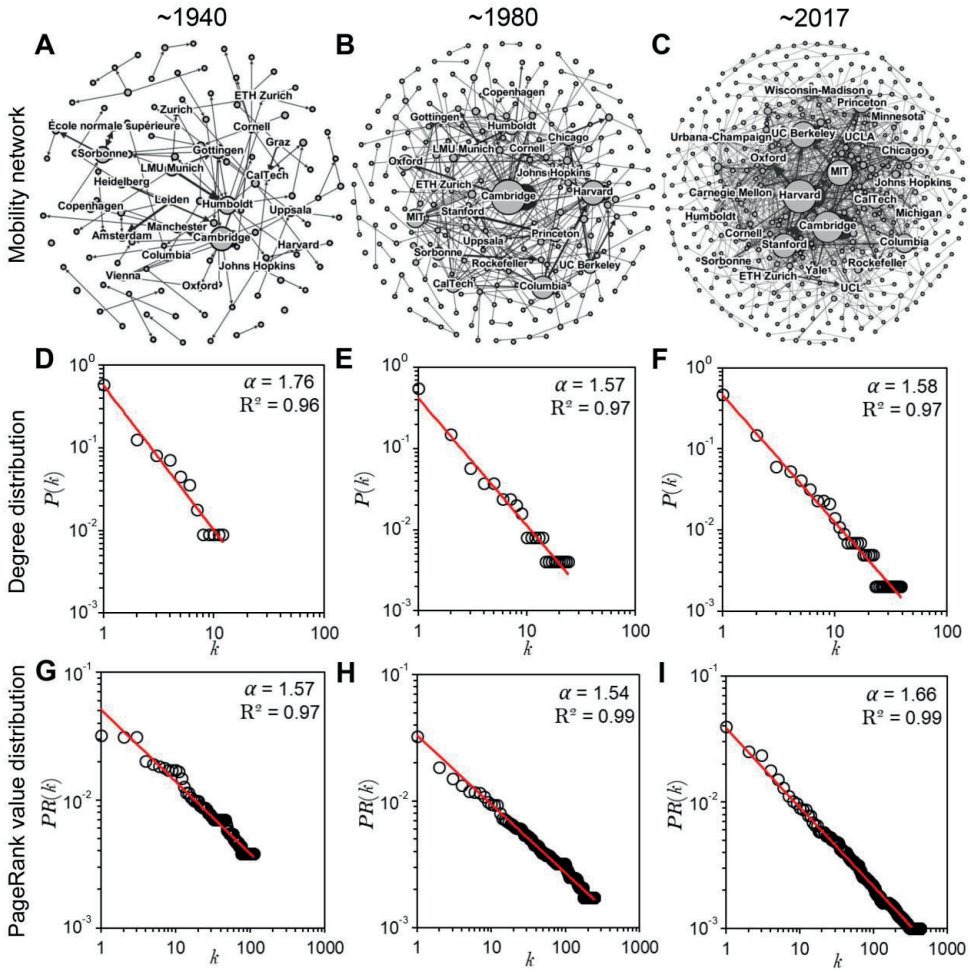


Figure 3. The laureate mobility networks at different time periods, from the year of 1901 to 1940 (A), 1980 (B), and 2017 (C), their degree distributions (D, E, and F), and their PageRank value distributions (G, H, and I).

On the other hand, the ranking of hub institutions has been changing over time, Table 3 lists the top five hub institutions at the time of 1940, 1980, and 2017. By the year of 1940 and 1980, University of Cambridge is the most hub institution, but in 2017 it changed to Harvard University.

Table 3. The top five hub institutions at different historical periods.

1940			1980			2017		
N(*k*)	*D*(*k*)	*PR*(*k*)	*N*(*k*)	*D*(*k*)	*PR*(*k*)	*N*(*k*)	*D*(*k*)	*PR*(*k*)
Cambridge	23	0.032	Cambridge	66	0.032	Harvard	148	0.039
Humboldt	16	0.029	Harvard	44	0.018	MIT	110	0.025
Cornell	6	0.027	Columbia	41	0.015	Cambridge	138	0.023
Gottingen	11	0.020	MIT	32	0.013	UC Berkeley	113	0.018
Manchester	6	0.019	Humboldt	24	0.011	Stanford	105	0.015

Level of institutional inequality

Figure 4 illustrates the Lorenz curves and Gini coefficients of laureate mobility networks at different historical periods and in different research fields. We find G = 0.46, 0.59, and 0.67 by the year of 1940, 1980, and 2017 respectively; G = 0.52, 0.55, and 0.57 in engineering sciences, social sciences, and natural sciences respectively. These number indicates the strong inequality across research fields and become expanding over time (56% and 76% of laureates' mobility were in 20% institutions by the year of 1940 and 2017 respectively).

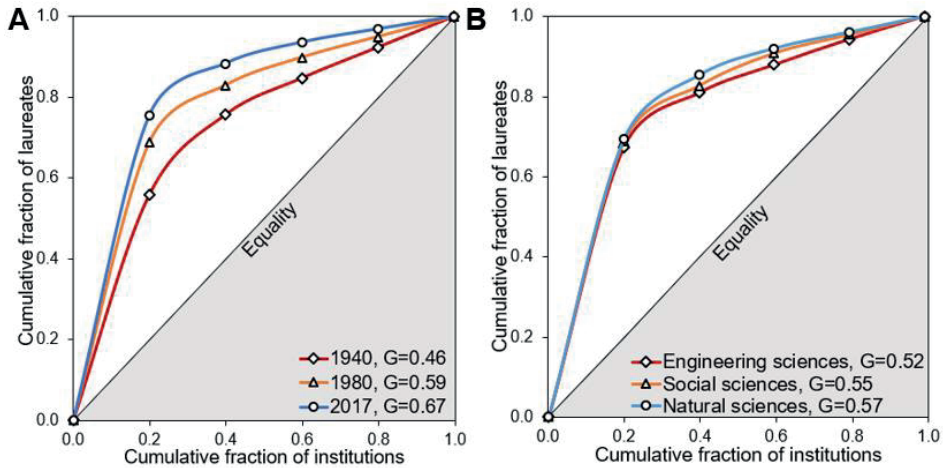


Figure 4. Lorenz curves and Gini coefficients showing the levels of institutional inequality in laureate mobility networks at different time periods (A) and in different research fields (B).

Furthermore, we note that the proportion of the degree of top one hub institutions to the cumulative degree of all institutions (d_1/d_{all}) is expanding over time. As shown in Figure 5, the proportion ranging from 0.15 to 0.2 before the year of 1960, yet dramatically increased to 0.25

in the year of 1980, and touched 0.34 in 2017. This proportion of 0.34 means that there are 34% of laureates either received their highest degree from Harvard University or worked at Harvard University when received the award.

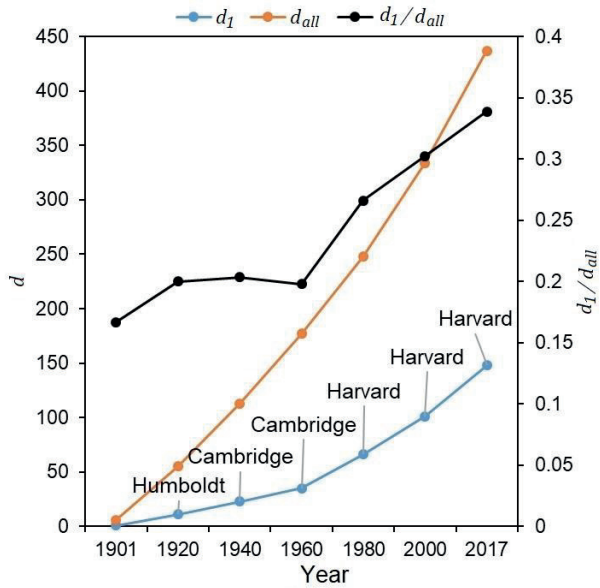


Figure 5. The cumulative degree of all institutions (d_{all}), and degree of the top one hub institutions (d_1) and its proportion (d_1/d_{all}) at different time periods.

Concluding remarks

In this paper, we examined the institutional hierarchy of mobility networks of prestigious awards laureates using tools derived from network theory, including structure detection and centrality analysis. Additionally, we apply measurements of Lorenz curve and Gini coefficient to observe the levels of institutional inequality. The study has a few limitations. The limited sample of our data may bias the measurement results of laureate mobility networks, as we only collected two relevant institutions (highest degree received institution and awarding time institution) of each laureate. Our further work would be to solid the observation by expanding the sample data of laureates and their mobility institutions.

So, is there an elite circulation phenomenon, or in other words, "new wine in old bottles"? Unfortunately, our findings did support a positive answer, from the evidence of scale-free properties, hub institutions over time, and the strong level of Gini coefficients in laureate mobility networks. These results also indicate that affiliation matters. More specifically, researchers at hub institutions (e.g. Harvard University, University of Cambridge, MIT, and UC Berkeley) may have better chances to receive prestigious awards than those at institutions located in the periphery of networks. Taking a step further, the institutional hierarchy of laureate mobility networks also reveals that Matthew effect in science has been aggravated over time, which may drive to a severe stratification of institution status in academia.

However, an unsolved problem of this study is that, from a social capital perspective, is the academic hierarchy an essential meritocratic structure that rewards researchers who have greater abilities? Or, does it suggest that this hierarchical structure reflects specific status

divisions? If so, what is the evolving mechanism behind this institutional stratification? These are more complicated questions that further studies should be devoted to answering them.

Finally, our findings also have some policy implications. As Gini coefficient has been successfully regarded as a measurement of inequality of income or wealth from economic sectors, which a Gini coefficient of 0.4~0.5 is generally regarded as the international warning level for dangerous levels of inequality in economic society (Tao, Wu & Li 2017). Thus, the academic sectors may also set a sort of "alarming level" of inequality in academia for social and public good.

References

Adams, W. (1968). *The Brain Drain*. New York: McMillan.

Arewa, O. B., Morriss, A. P., & Henerson, W. D. (2014). Enduring hierarchies in American legal education. *Indiana Law Journal, 89*(1), 941.

Barabási, A. L., & Albert, R. (1999). Emergence of scaling in random networks. *Science, 286*(5439), 509–512.

Barnett, G. A., Danowski, J. A., Feeley, T. H., & Stalker, J. (2010). Measuring quality in communication doctoral education using network analysis of faculty-hiring patterns. *Journal of Communication, 60*(2), 388-411.

Bernal, J. D. (1954). *Science in History*, London: CA Watts & Co.

Brin, S., & Page, L. (1998). The anatomy of a large-scale hypertextual web search engine. *Computer Networks and ISDN Systems, 30*(1–7), 107–117.

Burris, V. (2004). The academic caste system: Prestige hierarchies in PhD exchange networks. *American Sociological Review, 69*(2), 239-264.

Cantwell, B. (2011). Transnational mobility and international academic employment: Gatekeeping in an academic competition arena. *Minerva, 49*(4), 425-445.

Cañibano, C., Otamendi, F. J., & Solís, F. (2011). International temporary mobility of researchers: a cross-discipline study. *Scientometrics, 89*(2), 653-675.

Clauset, A., Arbesman, S., & Larremore, D. B. (2015). Systematic inequality and hierarchy in faculty hiring networks. *Science advances, 1*(1), e1400005.

Clauset, A., Shalizi, C. R., & Newman, M. E. (2009). Power-law distributions in empirical data. *SIAM review, 51*(4), 661–703.

Cruz-Castro, L., & Sanz-Menéndez, L. (2010). Mobility versus job stability: Assessing tenure and productivity outcomes. *Research Policy, 39*(1), 27-38.

Cole, J. R., & Cole, S. (1973). *Social stratification in science*. Chicago, IL: University of Chicago Press.

Coupé, T. (2003). Revealed performances: Worldwide rankings of economists and economics departments, 1990-2000. *Journal of the European Economic Association, 1*(6), 1309-1345.

Crane, D. (1969). Social structure in a group of scientists: A test of the "Invisible College" hypothesis. *American Sociological Review, 34*(3), 335-352.

Gaston, J. (1970). The reward system in British science. *American Sociological Review, 35*(4), 718-732.

Gastwirth, J. L. (1971). A general definition of the Lorenz curve. *Econometrica: Journal of the Econometric Society*, 1037-1039.

Gastwirth, J. L. (1972). The estimation of the Lorenz curve and Gini index. *The Review of Economics and Statistics*, 306-316.

Gleich, D. F. (2015). PageRank beyond the Web. *SIAM Review, 57*(3), 321-363.

Jiang, F., & Liu, N. (2018). The hierarchical status of international academic awards in social sciences. *Scientometrics, 117*(3), 2091-2115.

Katz, D. M., Gubler, J. R., Zelner, J., Bommarito, M. J., II, Provins, E., & Ingall, E. (2011). Reproduction of hierarchy? A social network analysis of the American Law Professoriate. *Journal of Legal Education, 61*(1), 76.

Mai, B., Liu, J., & González-Bailón, S. (2015). Network effects in the academic market: Mechanisms for hiring and placing PhDs in communication (2007-2014). *Journal of Communication, 65*(3), 558-583.

Melin, G. (2005). The dark side of mobility: negative experiences of doing a postdoc period abroad. *Research Evaluation, 14*(3), 229-237.
Mixon, F. G., Torgler, B., & Upadhyaya, K. P. (2017). Scholarly impact and the timing of major awards in economics. *Scientometrics, 112*(3), 1837-1852.
Morgan, C. A., Economou, J. D., Way, F. S., & Clauset, A. (2018). Prestige drives epistemic inequality in the diffusion of scientific ideas. *EPJ Data Science, 40*(7), 1-16.
Myers, S. A., Mucha, P. J., & Porter, M. A. (2011). Mathematical genealogy and department prestige. *Chaos: An Interdisciplinary Journal of Nonlinear Science, 21*(4), 041104.
Pareto, V. (1968). *The Rise and Fall of Elites: An Application of Theoretical Sociology*. New York: Arno Press.
Schlagberger, E. M., Bornmann, L., & Bauer, J. (2016). At what institutions did Nobel laureates do their prize-winning work? An analysis of biographical information on Nobel laureates from 1994 to 2014. *Scientometrics*, *109*(2), 723-767.
Tao, Y., Wu, X., & Li, C. (2017). Rawls' fairness, income distribution and alarming level of Gini coefficient. *Economics Discussion Papers, No 2017-67, Kiel Institute for the World Economy*. http://www.economics-ejournal.org/economics/discussionpapers/2017-67.
Wiggins, A., & Sawyer, S. (2012). Intellectual diversity and the faculty composition of iSchools. *Journal of the American Society for Information Science and Technology, 63*(1), 8-21.
Yuasa, M. (1962). Centre of scientific activity: Its shift from the 16th to the 20th century. *Japanese Studies in the History of Science, 1*, 57-75.
Zhang, Y., & Huang, M. M. (2018). Discussion of university performance-related pay incentive mechanism based on the analysis of Gini coefficient. *4th Annual International Conference on Management, Economics and Social Development (ICMESD 2018)*. Paris: Atlantis Press.
Zhao, H., & Jiang, G. (1985). Shifting of world's scientific center and scientists' social ages. *Scientometrics, 8*(1-2), 59-80.
Zheng, J., & Liu, N. (2015). Mapping of important international academic awards. *Scientometrics, 104*(3), 763-791.
Zhou, J., Zeng, A., Fan, Y., & Di, Z. (2018). Identifying important scholars via directed scientific collaboration networks. *Scientometrics, 114*(3), 1327-1343.
Zhu, Y., & Yan, E. (2017). Examining academic ranking and inequality in library and information science through faculty hiring networks. *Journal of Informetrics, 11*(2), 641-654.
Zubieta, A. F. (2009). Recognition and weak ties: is there a positive effect of postdoctoral position on academic performance and career development?. *Research Evaluation, 18*(2), 105-115.
Zuckerman, H. (1995). *Scientific elite: Nobel Laureates in the United States*. New Brunswick, NJ: Transaction Publishers.

The coverage of blogs and news in the three major altmetric data providers

José Luis Ortega[1]

[1] *jortega@iesa.csic.es*
Institute for Advanced Social Sciences, CSIC, Campo Santo de los Mártires, 7 14004 Córdoba (Spain)

Abstract

The objective of this study is to compare the coverage of the three main altmetric data providers (PlumX, Altmetric.com and Crossref Event Data) according to blog posts and news. The paper emphasizes the coverage the events that mention publications with the purpose of observing the media coverage and overlap between services. More than 100,000 random publications from Crossref were searched in the three providers. The link, title and source of the events that mention each document were retrieved. Results show that the number of publications mentioned in blogs (9.9%) and news (4.8%) is very low, being Altmetric.com the service with more documents mentioned by blogs (6%) and news (3.7%). Altmetric.com also covers more blogs (37.8%), while PlumX collects a large number of news media (36.5%). The overlap between the altmetric providers is rather low in publications (7.4%), events (0.5%) and sources (4.1%). The study concludes that the employment of several providers is necessary to undertake any reliable altmetric study.

Introduction

Altmetric data providers have gained great importance in the scholarly publication system because they capture and make public the mentions that receive academic documents in different web environments. Exploring an endless number of media, social networks, blogs, reference managers and web databases, these services find and index events that mention research outputs. Altmetric event is any action (citation, mention, saved, view, download, etc.) that occurs about a research papers on the Web. This information is acquiring great value for different stakeholders. Publishers esteem altmetric data because they can track the usage and valuation of the papers that they publish, observing who their audiences are and how they respond to their publications. Policy makers and funding agencies can appreciate the impact of the research that they fund in different social environments.

However, bibliometricians and altmetricians are the most demanding users because they utilize these tools as research instruments to analyse the meaning and behaviour of these metrics and to observe their involvement in evaluation activities. Altmetric studies are increasingly depending on these services to carry out their analyses. This dependence is based on the technical difficulty of tracking mentions from whole the Web and the need to establish agreements with third parties (Twitter, Facebook, etc.). This situation makes very difficult for researcher to undertake altmetric studies by their own means.

This fact is causing growing concern about the reliability and accuracy of these platforms as data providers because the validity of the results is strongly determined by the source used for the analysis. Several studies have addressed this issue, comparing the counts supplied by each platform in a sample of publications (Meschede and Siebenlist, 2018; Zahedi and Costas, 2018; Ortega, 2018). Results, in each case, have evidenced important differences in coverage of publications and metric counts. However, these results have been criticized because the comparison of total counts does not allow to study the overlap between providers (Wass, 2018). That is, comparing the total number of mentions in different platforms does not permit to know how many of them are repeated. Then, a fair comparison should be to identify all the posts that mention a publication and check if they are the same or not across different platforms. Unfortunately, it is not easy to obtain the original documents (tweets, blog posts, news, etc.) that mention a publication from these providers, which supposes an important problem to carry out studies about the overlap among altmetric providers.

This study tries to fill this gap, analysing the largest sample of publications (100,000 documents) ever used for comparing the three major altmetric providers (Altmetric.com, PlumX and Crossref Event Data (CED)). In this case, the study is limited to compare the coverage of blog posts and news with the aim of benchmarking not only the overlap between providers but also of checking the size of their lists of blogs and news outlets.

Literature review

In spite of the growing importance of data providers in the altmetric studies, the literature about the functionalities and working of these platforms is not very large. Adie and Roe (2013) were the first ones who detailed how Altmetric.com tracks the mention of papers on the Web. Trueger et al. (2015) made a critical review of the Altmetric Score, although, it was Gumpenberger et al. (2016) who expressed the strongest criticism about this indicator. In the case of PlumX, Champieux (2015) and Lindsay (2016) described the utilities of the service, while Wong and Vital (2017) analysed the implementation of the tool in a specific organization. However, no studies have been yet published about the functionality of CED mainly due to this product is still in beta.

Nevertheless, many other studies have analyzed the coverage of these services, describing the proportion of altmetric events in different samples. Thelwall et al. (2013) performed the earliest distribution of metrics in Altmetric.com, finding a greater proportion of papers mentioned in Twitter and Facebook. Robinson-García et al. (2014) also analyzed the coverage of this provider and they found that 87.1% of articles had at least one tweet and 64.8% one Mendeley reader. In a similar way, Bornmann (2014) explored a set of articles from Altmetric.com and he observed that 71% of articles were tweeted and a moderated proportion of documents were mentioned in Facebook (31%). More recently, Thelwall (2018) studied the coverage of Social Sciences, Arts and Humanities found a low prevalence (less than 12%, excepting tweets). According to PlumX, it worth mentioning the work of Torres-Salinas et al. (2017) about the collection of books. Their results showed that the distribution of events for books is rather different than for articles. Ortega (2018a) used PlumX data to track the life cycle of several altmetrics, observing that the most frequent ones were Mendeley readers and Tweets.

Specifically, several studies have focused on the blogs posts as altmetric indicators. Fausto et al. (2012) were the first ones to explore the relationship between blog posts and citations. Their result shown a positive correlation between post views and citations. Shema et al. (2014) found that articles receiving blog mentions close to their publication date receive more journal citations. These same authors (Shema et al., 2015) found that reviews and multidisciplinary top-tier journal articles were overrepresented in blog mentions. Jamali and Alimohammadi (2015) observed that discussion and criticism were the two main categories of motivations for citing articles in blogs. Contrarily, literature about the mention of research articles in news outlets is almost non-existent. Only, it worth to mention studies that describe the proportion of articles cited in news. According to Altmetric.com, Bornmann (2014) observed a moderated proportion of documents mentioned in news (13%), while Fraumann et al. (2015) found an important bias towards U. S. sites. MacLaughlin et al. (2018) studied the features that improve the popularity of research articles mentioned in news.

However, more papers have performed comparative studies between altmetric aggregators. Jobmann et al. (2014) were the first ones to compare the coverage and counts of ImpactStory, Altmetric Explorer, Plum Analytics and Webometric Analyst by research areas. Their results showed important divergences between services, being Plum Analytics the platform that better covered Mendeley and Facebook data, while Altmetric.com highlighted gathering blogs, news and CiteULike data. Zahedi et al. (2015) explored the consistency of data across Altmetric.com, Mendeley and Lagotto. They also detected significant differences, finding that Altmetric.com gathered more tweets, but it was less accurate collecting Mendeley readers. Baessa et al. (2015)

evaluated several altmetric services for their institutional repository and they recognized that Altmetric.com had a better coverage of blogs, news and government documents, while PlumX was most exhaustive covering different formats as books or reports. Kraker et al. (2015) studied the gathering of research data in Figshare, PlumX and ImpactStory. They observed that PlumX detected considerably more items in social media and higher scores than ImpactStory. Peters et al. (2016) extended their former study (Peters et al., 2015) with the inclusion of Altmetric.com. Their results confirmed that PlumX was the best provider for covering non published materials such as research data. Meschede and Siebenlist (2018) compared Altmetric.com and PlumX, finding that less than half of the publications analyzed were included in Altmetric.com, while PlumX covered almost the totality (99%). Zahedi and Costas (2018) performed the most exhaustive comparison between data providers, finding substantial differences in the metrics offered by these platforms. Bar-Ilan et al. (2018) compared Mendeley, Altmetric.com and PlumX in two time spots (2017 and 2018). Their results showed that the overlap between PlumX and Altmetric.com increased. Torres-Salinas et al. (2018) compared the coverage of books in Altmetric.com and PlumX and they concluded that they are rather complementary than comparable tools. Finally, Ortega (2018b) benchmarked the metrics counts of Altmetric.com, PlumX and CED, observing that Altmetric.com is the best aggregator of blog posts, news and tweets; PlumX of Mendeley readers; and CED of Wikipedia citations.

Objectives

The main objective of this study is to contrast, for first time, the coverage of the three main altmetric data providers according to blog posts and news. The aim is not only comparing the coverage of publications and number of counts, but also exploring the own events that mention the publications with the purpose of observing the media coverage and overlap between services. In particular, the following research questions were addressed:

- To what extent do the three providers cover the publications of the sample and what is the overlap between them?
- To what extent do the three providers capture blog posts and news that mention the publications of the sample and what is the overlap between them?
- What is the media coverage of the three providers and what is the overlap between them?

Methods

This study is particularly focused on blogs and news mentions. The reason for analysing only these metrics is that they do not come from a specific platform such as tweets (Twitter) or readers (Mendeley), but from a large number of media. This fact makes possible comparing the coverage of both documents and events, and the sources of those events. Another reason is that blogs and news are comparable metrics in the three providers.

Altmetric providers

PlumX: PlumX (plu.mx/plum/g/samples) is a provider of alternative metrics created in 2012 by Andrea Michalek and Michael Buschman from Plum Analytics. This product is addressed to the institutional market, offering altmetric counts of publications for particular institutions. PlumX is the aggregator that offers more metrics, including citation and usage metrics (i.e. Views and Downloads). It covers more than 52.6 million of artifacts, being the largest altmetric aggregator (Plum Analytics, 2018). In 2017, Plum Analytics was acquired by Elsevier (www.elsevier.com), tracking now the online presence of any article indexed in Scopus database (Elsevier, 2017). This agreement with Elsevier also caused that PlumX used Newsflo (an Elsevier company) as news data provider (Allen, 2017). On the other hand, PlumX also

includes blog mentions from ACI Scholarly Blog Index (ACI, 2016). However, PlumX does not provide any information about the number of blogs and media covered.

Altmetric.com: It was the first altmetric provider and it was initiated in 2011 by Euan Adie, with the support of Digital Science (www.altmetric.com). Unlike PlumX, Altmetric.com is centered in the publishing world, signing agreements with publisher houses to monitor the altmetric impact of their publications. This information is accessible through a public API. Today, Altmetric.com tracks the social impact of close to 9 million of research papers (Altmetric.com, 2018). Altmetric.com monitors around 14,000 blogs (Altmetric.com, 2019), although it does not make available the list of sources. According to news, Altmetric.com collects a list of 2,900 news outlets (Altmetric.com, 2019). This list is publicly available on the site.

Crossref Event data (CED): CED is the youngest service. Created in 2016, it is still in beta (www.crossref.org/services/event-data). Unlike the previous ones, CED is not a commercial site and it provides free access to their data through a public API. Another important difference is that it does not provide metrics, but it only displays information about each altmetric event linked to a DOI identifier. For instance, it shows the information about the mention of an article on Twitter (date, user, tweet, etc.), but it does not show a count of the number of tweets. For that reason, CED's data would have to be processed to be comparable with the other services. CED does not distinguish between blogs and news. It defines three categories, *wordpressdotcom*, *web* and *newsfeed*s to group blogs and news. In addition, category *reddit-links* includes links from Reddit that point external sources such as blogs and news. A manual inspection evidenced that the media and blogs are equally classified as *web* or as *newsfeed*, and sometimes in both categories at the same time. Due to this, the distinction between blog and news is based on the matching with the other data providers. In the case of mentions that do not match, then a manual classification was done.

Data extraction

This study aims to compare the coverage of blogs and news mentions by the three major altmetric providers. A random sample of 100,529 DOIs from Crossref were extracted to detect the number of publications covered by these aggregators. These publications were obtained from Crossref API with the conditions of being journal articles and published from 2012 (https://api.crossref.org/works?sample=100&filter=type:journal-article,from-pub-date:2012-01-01). 2012 was elected because is the time window sufficiently broad to capture the impact of the sample in blogs and news.

Next, this list was searched in the three data providers. In the case of Almetric.com, Altmetric ID was obtained from the Altmetric API (api.altmetric.com/v1/doi/), and then it was used to extract data about blogs and news directly from the web site (www.altmetric.com/details/). This is due to the API only shows counts and not the links and content of these mentions. In the case of PlumX, DOIs were searched in the web site of PlumX (plu.mx/plum/a/?doi=). Finally, information of CED was extracted from the API (query.eventdata.crossref.org/events?filter=obj-id:). In the three cases, several SQL scripts were written to scrape the data from websites and APIs. This process was performed during the second fortnight of August 2018.

Results

Publications

Our first objective is to know the overlap of blogs and news between the three data providers. Table 1 shows the number and proportion of articles indexed in these providers, the number

and proportion of articles with any mention, and the relationship between blogs and news in each platform.

Table 1. Distribution of articles in the three major data providers

	Altmetric.com	*PlumX*	*CED*	*Sample*
Articles	34,583	99,697	100,116	100,529
% Sample	34.4%	99.2%	99.6%	11,728
Mentioned articles	7,457	7,274	1,698	11,7%
% Sample	7.4%	7.2%	1.7%	
% Provider	21.6%	7.3%	1.7%	9,990
Blogs	5,995	5,741	1,222	11,728
% Sample	6.0%	5.7%	1,222	9,990
% Provider	17.3%	5.8%	1.2%	9.9%
News	3,712	2,784	1.2%	
% Sample	3.7%	2.8%	662	4,862
% Provider	10.7%	2.8%	0.7%	4.8%

CED is the service that indexes more journal articles (99.6%) because Crossref provides its database of articles to CED. The remaining .4% could be to technical errors during the data extraction process. PlumX is the second one in coverage with 99.2% of the sample, which it makes clear its strong capability to gather publications. Altmetric.com, however, only covers the 34% of the sample, which confirms the reduced size of this tool (Meschede and Siebenlist, 2018; Ortega, 2018). Nevertheless, not all the publications indexed in a provider have mentions. Concretely, regarding to blogs and news mentions, Altmetric.com is the service that have more articles mentioned in blogs and news (7.4%), followed shortly by PlumX (7.2%) and CED (1.7%). These results show that PlumX and Altmetric.com, independently of their sizes, gather a similar proportion of articles discussed in media. Other important result is that CED, in spite of its important coverage of publications, captures only a small fraction of events. According to the number of articles mentioned in blogs and news, PlumX and Altmetric.com show a similar proportion with regard to blogs (A=6%; P=5.7%), while Altmetric.com has a bit more articles mentioned in news media (A=3.7%; P=2.8%). CED shows a similar proportion with 1.2% of blogs and 0.7% of news.

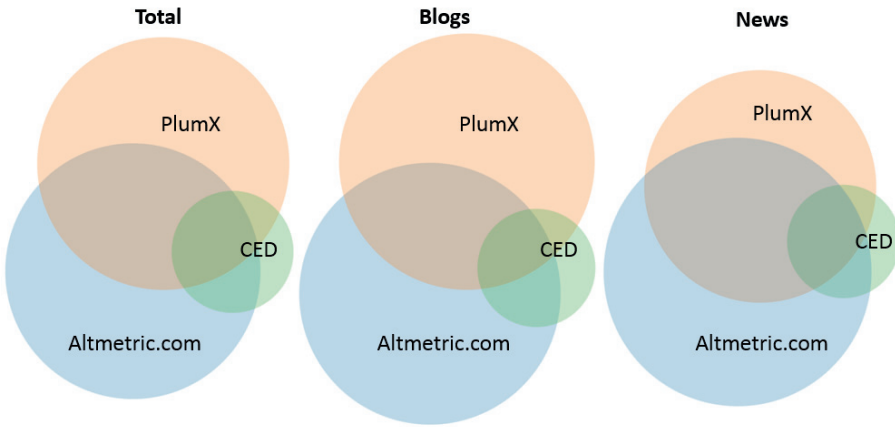


Figure 1. Venn diagram about the distribution and overlap of publications in the three major data providers.

Figure 1 shows the overlap of publications mentioned in the three altmetric providers according to news and blogs. The picture shows that there is low overlap between them (7.4%), being greater for news (7.9%) than for blogs (4.8%). The highest overlap is found between Altmetric.com and PlumX (28.8%), being more significant for news (36.8%) than for blogs (21.1%). These differences between blogs and news could be to the number of media is lower than the number of blogs, which could cause a higher overlap.

Events

Another way, although much more complex, to benchmark altmetric providers is to compare the number of events. URL of each event was used as identifier because different news and blog posts can have the same title. However, the URL matching could be problematic because both Altmetric.com and PlumX use third parties to provide links about blogs and news. In the case of Altmetric.com, 39.7% of the news outlets links come from Moreover.com (now Lexis-Nexis), while 5.2% of the blog posts links in PlumX are provided by ACI Scholarly Blog Index (now ProQuest). These services use link resolvers, and therefore the real URL is hidden. This problem gets worse when both services are now disappeared, and the links, in some cases, do not work properly. In the case of Altmetric.com, 8,794 (74.1%) links were resolved, while any link from ACI in PlumX could be resolved.

In addition, it was detected that some mentions do not actually come from news and blogs. Instead, they are citations from other research publications. This problem is not relevant in Altmetric.com because they just represent 0.8%. The real problem occurs in PlumX where 23.7% of the mentions are in fact bibliographic citations, mainly from Hindawi (17.9%) and OMICS Publishing Group (3.1%). These mistakes were detected more frequently in blogs (93.5%) than in news (6.5%). These citations were removed to do a fair comparison.

Another problem is that Altmetric.com does not provide full information about some events. In certain publications, Altmetric.com only shows the first four events while keeping the remaining ones hidden. The size of this gaps can be observed if the total number of events and the sum of counts are compared (Table 2). Notice that many events can mention more than one article, and then the number of real events has to be less than the cumulative count. This difference is significant in the number of news, where Altmetric.com only shows a 62% of the total counts. However, the percentage of blog posts (83.7%) could be considered acceptable. In PlumX, the percentage of blogs (27.9%) and news (90.2%) are caused by duplicated events and the omission of bibliographic citations aforementioned. There is not differences in CED because this provider does not show counts, but raw data.

Table 2. Distribution of events in the three major data providers

	Altmetric.com	*PlumX*	*CED*	*Sample*
Events	25,395	18,405	2,529	45,526
% Sample	55.8%	40.4%	5.5%	
Blogs	13,604 (83.7%)	3,216 (27.9%)	1,530	17,136
Count	16,242	11,537	1,530	
% Sample	29.9%	7.1%	3.4%	37.6%
% Provider	53.6%	17.5%	60.5%	
News	15,637 (61.7%)	15,288 (90.2%)	1,000	29,271
Count	25,321	16,949	1,000	
% Sample	34.3%	33.6%	2.2%	64.3%
% Provider	61.6%	83.1%	39.5%	

According to the distribution of events across the three providers, Table 2 shows the number and percentage of blog posts and news in Altmetric.com, PlumX and CED. In general, results show that Altmetric.com (55.8%) collects more events than PlumX (40.4%), while CED only achieves to capture a small amount of events (5.5%). However, each provider shows different proportions between blog posts and news. Altmetric.com is the source that includes more blog posts in the sample (29.9%), followed by PlumX (7.1%) and CED (3.4%). Whereas, Altmetric.com (34.3%) and PlumX (33.6%) captures a similar proportion of news, and much more than CED (2.2%). With regard to the proportion of blogs and news in each provider, the sample shows a higher proportion of news (64.3%) than blog posts (37.6%). PlumX is the service that presents a more disproportionate distribution (blogs=17.5%, news=83.1%), followed by CED (blogs=60.5%, news=39.5%). Altmetric.com is the service that shows a more equilibrate proportion (blogs=53.6%, news=61.6%).

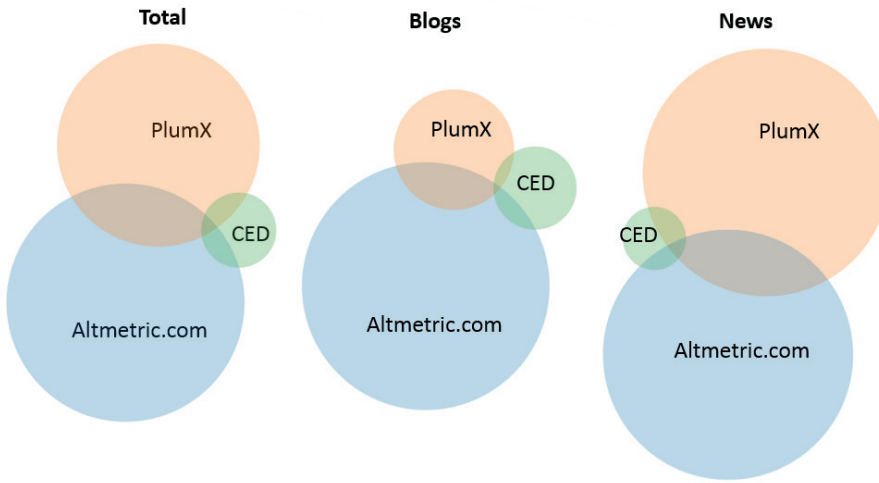


Figure 2. Venn diagram about the distribution and overlap of events in the three major data providers.

Figure 2 shows the overlap of blogs and news mentions between the three altmetric providers. Unlike Figure 1, the overlap of events is even lower than the overlap of publications. In general, only 0.5% of the events are simultaneously gathered by Altmetric.com, PlumX and CED, being the greatest overlap between Altmetric.com and PlumX (7.7%). According to blog posts, the overlap is diminished with only 0.3%, where Altmetric.com and PlumX gather together 5.5% of the blog posts. In the case of news, the overlap is slight higher (0.4%), being again Altmetric.com and PlumX the services that show the highest overlap (7.7%). This very low overlap among events evidences that the mention of articles in blogs and news is infrequent and they appear in a very varied range of sources that are not completely covered by all the altmetric providers together.

Sources

A third way to compare the altmetric providers is analysing the number of distinct sources that publish the blog posts and the news. Web domains and names of the sources were revised to merged duplicated sources. For example, blogs that change from web servers (botany.one, aobblog.com) or that have several domains (academiclifeinem.com, academiclifeinem.blogspot.com) were merged. Blogs hosted in the same platform such as *AGU*

Blogosphere (blogs.agu.org) and *LSE Blogs* (blogs.lse.ac.uk) were distinguished. News media with different languages editions (for example, *CNN* and *CNN en Español*) were differentiated as well, because it is assumed that the content is different in those editions.

Table 3. Distribution of sources in the three major data providers

	Altmetric.com	*PlumX*	*CED*	*Sample*
Sources	3,856	3,255	1,263	6,837
% Sample	56.4%	47.6%	18.5%	
Blogs	2,582	860	960	3,810
% Sample	37.8%	12.6%	14.0%	55.7%
% Provider	67.0%	26.4%	76.0%	
News	1,408	2,496	310	3,387
% Sample	20.6%	36.5%	4.5%	49.5%
% Provider	36.5%	76.7%	24.5%	

Table 3 details the number and percentage of sources found in the sample. Overall, 6,837 sources were detected. Altmetric.com contains 3,856 (56.4%) different sources, followed by PlumX with 3,255 (47.6%) and CED far away with 1,263 (18.5%). In general, the proportion of blogs 3,810 (55.7%) and news outlets 3,387 (49.5%) is rather equilibrated. However, this balanced distribution is broken when blogs and news are distinguished. According to blogs, Altmetric.com includes 2,582 (37.8%) sources, more than double that PlumX with 860 (12.6%) and CED with 960 (14%). This result emphasizes the good coverage of blogs by Altmetric.com. However, if the proportion of news media is observed, PlumX includes 2,496 (36.5%) sources, whereas Altmetric.com indexes 1,408 (20.6%) media and CED only 310 ones (4.5%). These differences could be due to the special coverage of the US local media by PlumX (Elsevier, 2018).

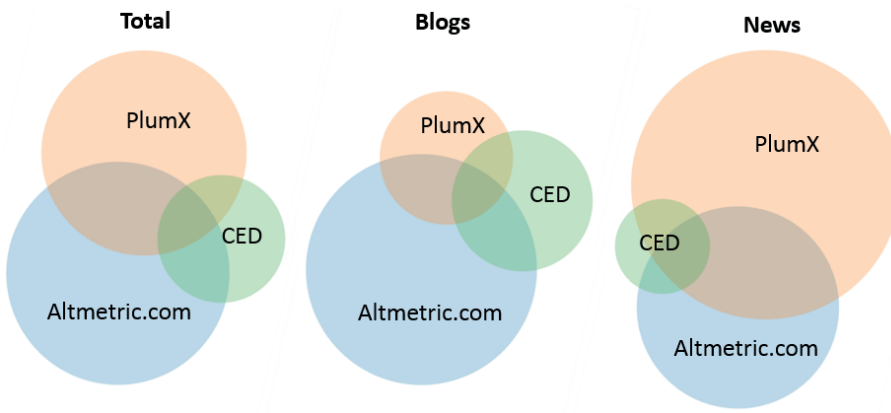


Figure 3. Venn diagram about the distribution and overlap of sources in the three major data providers.

Figure 3 shows the overlap of blogs and news sources between the three altmetric providers. Overall, the overlap between the three providers still is low (4.1%), being the couple Altmetric-PlumX which has the largest overlap (16%) and PlumX and CED are the aggregators that shares less sources (5.6%). According to blogs, the overlap between the three providers is only 1.6%, being again Altmetric.com-PlumX the pair that is more overlapped (5.7%) and PlumX and CED the services that share less sources (3.4%). News presents a little more overlap than blogs

(1.8%), being larger between Altmetric.com and PlumX (9.7%). However, CED and Altmetric.com are the providers that have less news in common (2.3%).

Discussion

The results on the detailed coverage of Altmetric.com, PlumX and CED of blog posts and news have showed a great disparity between the indexation of articles, the capture of events and the coverage of sources. PlumX and CED index almost all the articles of the sample (99%), while Almetric.com only covers 34%. The cause of this low coverage is unknown and it could be due to publisher's agreements or that Altmetric.com only indexes articles that have some social event. Precisely, this last reason could be the most probable one, as Altmetric.com is the platform that has more articles mentioned in blogs and news media (7.4%), followed by PlumX (7.2%) and CED (1.7%). Results show that there is almost the double of articles cited in blogs (9.9%) than in news (4.8%), which suggests that the blog post format is more suitable for the discussion of new academic results (Shema et al., 2015; Ritson, 2016), while traditional news format is used more to disseminate outstanding advances (Bubela and Caulfield, 2004; Suleski and Ibaraki, 2010). In this sense, Altmetric.com is the service that more articles indexes in both type of events (blogs=6%, news=3.7%), followed by PlumX (blogs=5.7%, news=2.8%), and CED (blogs=1.2%, news=0.7%).

One of the most relevant facts in this study is that it is not based on counts but on events. This element had made possible to compare data providers according to their coverage of blogs posts and news. In this sense, it is interesting to emphasize that Altmetric.com is the service that captures more events (55.8%) and from more different sources (56.4%), with a well-adjusted distribution between blogs (29.9%) and news (34.3%). However, PlumX presents a more imbalanced proportion, with a very low percentage of blog posts (7.1%) and sources (12.6%) opposite to a good coverage of news (33.6%) and news outlets (36.5%). This last percentage is even above of Altmetric.com (20.6%). Two elements explain this biased distribution. First, PlumX has a deficient coverage of blogs because it includes research articles as blogs, which causes a miscount of blog posts. Second, PlumX shows a special coverage of local US media (TV and radio stations), which increases the mentions of articles from news (Elsevier, 2018). Finally, CED remains in a secondary role because it shows much lower figures than the two before services. For example, CED only covers 5.5% of the events, being 3.4% of blog posts and 2.2% of news. It is only remarkable that CED indexes more blog sources (14%) than PlumX (12.6%).

These important differences in the coverage of blogs and news is also reflected in the low overlap between services. Only 7.9% of the articles in the sample are indexed by the three data providers, being the highest overlap between PlumX and Altmetric.com (28.8%). This overlap is even more reduced when the events and sources are observed. Thus, 0.5% of the events and 4.1% of the sources are simultaneously gathered by the three platforms, being again the greatest overlap between Altmetric.com and PlumX (events=7.7%; sources=16%). This low overlap is consequence of the low number of scholarly results being commented in blogs and news (11.7%), which cause a great spreading of the mentions in a wide range of sources. In fact, the study gathers less than 4.000 sources for Altmetric.com and 3.200 for PlumX, a very low number if they are compared with the official figures (14,000 blogs and 2,900 news outlets for Altmetric.com and 10,000 blogs and 55,000 news sources for PlumX). The low correlation in News (r=0.11) found by Meschede and Siebenlist (2018) confirms the great disparity between Altmetric and PlumX. However, all these disparities do not do anything but confirm that only one provider is not enough to observe the social impact of a publication and it is necessary the employment of several aggregators to undertake any reliable altmetric study.

However, this study goes beyond the event counts and, for first time, compares the content of the blog posts and news. This approach has disclosed import inconsistence in the count of these

metrics. The distinction between blogs and news is not explained and many sources are classified in both groups. This happens in 3.5% of sources in Altmetric.com and 3.2% in PlumX. Although this percentage is low, it is indeed significant and it could influence the final count of blogs and news. However, without a doubt, the most important problem is caused by the inclusion of bibliographic citations as blog mentions. This problem is insignificant in Altmetric.com (0.8%), but in PlumX is a serious issue because 23.7% of the blog mentions are, in fact, bibliographic citations. This result questions the reliability and accuracy of PlumX as data provider, concretely with regard to blogs.

The way in which the data were captured leads us to consider some important limitations when the results are interpreted. The most important limitation is that Altmetric.com, concretely Altmetric Explorer, does not show, in some cases, all the mentions that an article receives. This problem is especially significant in news where approximately 38% of the events were not retrieved. This problem can distort the results about coverage and overlapping, mainly, in the cases of events and sources. Another limitation is that many of the mentions come from disappeared services (Moreover.com and ACI Scholarly Blog Index), which caused that many of the events (5.2% of the blog posts in PlumX and 19.8% of the news in Altmetric.com) could not be verified and therefore compared with the other services. This technical problem would be able to influence the before figures about overlapping. However, this problem introduces the issue of the ability of these services to be audited by an independent organization that verified the counts that they publish (NISO, 2016). Anyways, future studies that test the coverage and overlapping between altmetric providers are welcome.

Conclusions

Several conclusions can be drawn from the results. First, PlumX and CED are the services that index more publications, reaching almost the totality of the sample. However, Altmetric.com is the service that gathers more documents mentioned in blogs and news (7.4%), closely followed by PlumX (7.2%) and CED (1.7%). The overlap between the altmetric providers is low (7.4%), being the greatest one between Altmetric.com and PlumX (28.8%).

Second, Altmetric.com is the platform that more blog and news events captures (55.8%), followed by PlumX (40.4%) and CED (5.5%). The proportion between blogs and news is balanced in Altmetric.com (blogs=29.9%; news=34.3%) and CED (blogs=3.4%; news=2.2%). However, PlumX has an important gap in the coverage of blog posts (7.1%) due to the miscounting of bibliographic citations as blog mentions. The overlap between platforms according to events is lower than according publications (0.5%), being the greatest for the couple Altmetric.com-PlumX (7.7%).

Third, according the sources of the events, Altmetric.com (56.4%) collects more distinct sources than PlumX (47.6%) and CED (18.5%). Again, Altmetric.com covers more blogs (37.8%), due to its exhaustive list of blogs. Whereas, PlumX highlights covering news outlets (36.5%), caused by the special coverage of local US media by its news provider, Newsflo. The overlap between the three services is again low (4.1%) and it makes evident the great amount of media that mention research outputs and the difficulty of gathering this information.

References

ACI (2016). ACI Scholarly Blogs Now Included in PlumX. Retrieved January 2, 2019 from: https://aci.info/2016/09/27/aci-scholarly-blogs-now-included-in-plumx/

Adie, E. & Roe, W. (2013). Altmetric: enriching scholarly content with article-level discussion and metrics. *Learned Publishing*, *26*(1), 11-17.

Allen, P. (2017). World-Wide News Coverage in PlumX. Retrieved January 2, 2019 from: https://plumanalytics.com/world-wide-news-coverage-plumx/

Altmetric.com (2018). How it works. Retrieved January 2, 2019 from: https://www.altmetric.com/about-our-data/how-it-works/

Altmetric.com (2019). What outputs and sources does Altmetric track? Altmetric.com Support. Retrieved May 6, 2019 from: https://help.altmetric.com/support/solutions/articles/6000060968-what-outputs-and-sources-does-altmetric-track-

Baessa, M., Lery, T., Grenz, D. & Vijayakumar, J. K. (2015). Connecting the pieces: Using ORCIDs to improve research impact and repositories. *F1000Research*, *4,* 195.

Bar-Ilan, J., Halevi, G. & Milojevic, S. (2018). Differences between Altmetric Data Sources–A Case Study. In *23rd International Conference on Science and Technology Indicators (STI 2018)* (pp. 699-702). Leiden: CWTS.

Bornmann, L. (2014). Validity of altmetrics data for measuring societal impact: A study using data from Altmetric and F1000Prime. *Journal of Informetrics*, *8*(4), 935-950.

Bubela, T. M. & Caulfield, T. A. (2004). Do the print media "hype" genetic research? A comparison of newspaper stories and peer-reviewed research papers. *Canadian Medical Association Journal*, *170*(9), 1399-1407.

Cave, R. (2013). Overview of the Altmetrics Landscape. In *Proceedings of the Charleston Library Conference*. Retrieved January 2, 2019 from: https://core.ac.uk/download/pdf/12982803.pdf

Chamberlain, S. (2013). Consuming article-level metrics: Observations and lessons from comparing aggregator provider. *Information Standards Quarterly*, 25(2), 4-13.

Elsevier. (2017). Elsevier Acquires Leading 'Altmetrics' Provider Plum Analytics. Retrieved January 2, 2019 from: https://www.elsevier.com/about/press-releases/corporate/elsevier-acquires-leading-altmetrics-provider-plum-analytics

Elsevier. (2018). Newsflo > Coverage and sources. Retrieved January 2, 2019 from: https://www.elsevier.com/solutions/newsflo/sources

Fraumann, G., Zahedi, Z. & Costas, C. R. (2015). What do we know about Altmetric.com sources? A study of the top 200 blogs and news sites mentioning scholarly output. In *The 2015 Altmetrics Workshop*. Retrieved January 2, 2019 from: https://altmetrics.org/wp-content/uploads/2015/09/altmetrics15_paper_19.pdf

Gumpenberger, C., Glänzel, W. & Gorraiz, J. (2016). The ecstasy and the agony of the altmetric score. *Scientometrics*, *108*(2), 977-982.

Jamali, H. R. & Alimohammadi, D. (2015). Blog Citations as Indicators of the Societal Impact of Research. *International Journal of Knowledge Content Development & Technology*, *5*(1), 15-32.

Jobmann, A., Hoffmann, C.P., Künne, S., Peters, I., Schmitz, J. & Wollnik-Korn, G. (2014). Altmetrics for large, multidisciplinary research groups: Comparison of current tools. *Bibliometrie – Praxis und Forschung, 3.* Retrieved January 2, 2019 from: http://www.bibliometrie-pf.de/article/viewFile/205/258

Kraker, P., Lex, E., Gorraiz, J., Gumpenberger, C. & Peters, I. (2015). Research data explored II: The anatomy and reception of figshare. *arXiv preprint arXiv:1503.01298*. Retrieved January 2, 2019 from: https://arxiv.org/ftp/arxiv/papers/1503/1503.01298.pdf

Lindsay, J. M. (2016). PlumX from Plum Analytics: Not Just Altmetrics. *Journal of Electronic Resources in Medical Libraries*, *13*(1), 8-17.

MacLaughlin, A., Wihbey, J. & Smith, D. A. (2018). Predicting News Coverage of Scientific Articles. In *12th International AAAI Conference on Web and Social Media (*ICWSM 2018) (pp. 191-200). Palo Alto, CA: AAAI Press

Meschede, C. & Siebenlist, T. (2018). Cross-metric compatability and inconsistencies of altmetrics. *Scientometrics*, *115*(1), 283-297.

NISO (2016). Outputs of the NISO Alternative Assessment Metrics Project. NISO RP-25-2016. Retrieved January 2, 2019 from: https://groups.niso.org/apps/group_public/download.php/17091/NISO%20RP-25-2016%20Outputs%20of%20the%20NISO%20Alternative%20Assessment%20Project.pdf

Ortega, J. L. (2018a). The Life Cycle of Altmetric Impact: A Longitudinal Study of Six Metrics from PlumX, *Journal of Informetrics*, *12*(3), 579-589.

Ortega, J. L. (2018b). Reliability and accuracy of altmetric providers: a comparison among Altmetric.com, PlumX and Crossref Event Data. *Scientometrics, 116*(3), 2123-2138.

Peters, I., Kraker, P., Lex, E., Gumpenberger, C. & Gorraiz, J. (2015). Research data explored: Citations versus altmetrics. In A.A. Salah, Y. Tonta, A.A. Akdag Salah, C. Sugimoto, & U. Al (Eds.),

Proceedings of the 15th International Conference on Scientometrics and Informetrics, Istanbul, Turkey (pp. 172–183)

Peters, I., Kraker, P., Lex, E., Gumpenberger, C. & Gorraiz, J. (2016). Research data explored: an extended analysis of citations and altmetrics. *Scientometrics*, *107*(2), 723-744.

Plum Analytics (2018). Coverage: Expanding the World of Altmetrics. Retrieved January 2, 2019 from: https://plumanalytics.com/learn/about-metrics/coverage/

Ritson, S. (2016). 'Crackpots' and 'active researchers': The controversy over links between arXiv and the scientific blogosphere. *Social studies of science*, *46*(4), 607-628.

Shema, H., Bar-Ilan, J. & Thelwall, M. (2014). Do blog citations correlate with a higher number of future citations? Research blogs as a potential source for alternative metrics. *Journal of the Association for Information Science and Technology*, *65*(5), 1018-1027.

Shema, H., Bar-Ilan, J. & Thelwall, M. (2015). How is research blogged? A content analysis approach. *Journal of the Association for Information Science and Technology*, *66*(6), 1136-1149.

Suleski, J. & Ibaraki, M. (2010). Scientists are talking, but mostly to each other: a quantitative analysis of research represented in mass media. *Public Understanding of Science*, *19*(1), 115-125.

Robinson-García, N., Torres-Salinas, D., Zahedi, Z. & Costas, R. (2014). Nuevos datos, nuevas posibilidades: Revelando el interior de Altmetric.com. *El Profesional de la Information*, *23*(4), 359-366.

Thelwall, M. (2018). Altmetric Prevalence in the Social Sciences, Arts and Humanities: Where are the Online Discussions?. *Journal of Altmetrics*, *1*(1). Retrieved January 2, 2019 from: https://www.journalofaltmetrics.org/articles/10.29024/joa.6/

Torres-Salinas, D., Gumpenberger, C. & Gorraiz, J. (2017). PlumX as a potential tool to assess the macroscopic multidimensional impact of books. *Frontiers in research metrics and analytics*, *2*, 5.

Torres-Salinas, D., Gorraiz, J. & Robinson-Garcia, N. (2018). The insoluble problems of books: What does Altmetric. com have to offer?. *Aslib Journal of Information Management*, *70*(6), 691-707.

Trueger, N. S., Thoma, B., Hsu, C. H., Sullivan, D., Peters, L. & Lin, M. (2015). The altmetric score: a new measure for article-level dissemination and impact. *Annals of emergency medicine*, *66*(5), 549-553.

Wass, J. (2018). Some thoughts on 'General discussion of data quality challenges in social media metrics'. *Joe's Blog*. Retrieved January 2, 2019 from: https://blog.afandian.com/2018/05/zahedi-costas-altmetrics/

Wong, E. Y. & Vital, S. M. (2017). PlumX: a tool to showcase academic profile and distinction. *Digital Library Perspectives*, *33*(4), 305-313.

Zahedi, Z., Fenner, M. & Costas, R. (2015). Consistency among altmetrics data provider/aggregators: What are the challenges?. In *The 2015 Altmetrics Workshop*. Retrieved January 2, 2019 from: https://altmetrics.org/altmetrics15/zahedi/

Zahedi, Z. & Costas, R. (2018). General discussion of data quality challenges in social media metrics: Extensive comparison of four major altmetric data aggregators. *PloS one*, *13*(5), e0197326.

Using a keyword extraction pipeline to understand concepts in future work sections of research papers

Kai Li[1] and Erjia Yan[2]

[1] *kl696@drexel.edu*
Drexel University, Philadelphia PA, 19104, United States

[2] *erjia.yan@drexel.edu*
Drexel University, Philadelphia PA, 19104, United States

Abstract

This paper presents a methodological framework, based on natural language processing (NLP) techniques, for identifying *future work* sentences in full-text scientific papers and extracting keywords from these sentences. We conduct a baseline test to evaluate the method's effectiveness and use full-text papers from *Science Advances* for a proof of concept. Our results suggest that there are significant domain differences in the extent to which keywords in the future work section match those in the title and abstract texts. This is, to our knowledge, the first empirical examination of future work statements in scientific articles. Our framework could contribute greatly to quantitative and predictive studies of science by introducing a new, more future-oriented data source, with potentially significant implications in both theory and practice. Moreover, our proof-of-concept study offers the first piece of evidence about the future work statement as a subgenre of scholarly writing. This evidence may inspire future scientometric studies, leading to a better understanding of the conceptual connections among papers published at different times.

Introduction

Scientometric analyses are highly retrospective, largely because of the nature of citation data. At the same time, many of these studies aim to understand the future: for example, by predicting emerging research topics and identifying "sleeping beauties" in the research literature[1]. As a result, a marked gap persists between the data available to scientometricians and their goal of predicting future research. For a long time, scientometricians only have had access to publication metadata—which, though useful for portraying the research landscape and mapping scientific domains, are insufficient to yield concise, contextualized insights for the forecasting of future research.

The *future work* sections of scientific papers could serve as an important data source in this new research agenda. As direct descriptions of what the authors deem important to be achieved in the future by themselves or others, these statements are a potential source of valuable forecasting data to inform science policymaking and planning. Combining existing methods from scientometrics and linguistics with advanced natural language processing (NLP) techniques, research into the future work section will provide a significant alternative for quantitative studies of science—transforming the metadata-based, retrospective mode of inquiry into a full-text-enabled, predictive one.

The promising role of future work statements in knowledge studies is nevertheless restricted by the fact that the future work section, as a subgenre, has seldom been examined by researchers. This is in direct contrast with other paper sections within the dominant IMRaD format (Huth, 1987; Sollaci & Pereira, 2004): the most notable examples include Abstract (Cross & Oppenheim, 2006; Dos Santos, 1996; Pho, 2008; Samraj, 2005) and Introduction (Gledhill, 2000; Samraj, 2002, 2005). Part of the difference between future work and these other scientific subgenres is that the former does not have clear-cut communicative functions and is not

[1] It should be noted that most of these studies are designed to investigate emerging topics and sleeping beauties that happened in the past (Ke, Ferrara, Radicchi, & Flammini, 2015; van Raan, 2004).

connected to any substantial reward (Teufel, 2017), which is only the first barrier to overcome before the overall research plan can be fulfilled.

In this paper, we present an NLP pipeline to identify future work statements in scientific papers and extract keywords from these sentences. This new tool aims to addresses the aforementioned gap between the potential predictive value of future work sections and the lack of research about this subgenre of scientific writing. We conduct a baseline test to evaluate the tool's precision and report the results of a proof of concept using full-text articles from the journal of *Science Advances*.

The present study is the first empirical effort to understand the linguistic and conceptual attributes of future work statements. Methodologically, our framework can be applied to any full-text corpora to better understand future work statements on a larger scale. It is, to our knowledge, an unprecedented effort to bring the future work section within the purview of quantitative and predictive studies of science.

Moreover, we hope to offer the first piece of empirical evidence as to the nature of conceptual connections between scientific papers and their future work statements. To pursue this question, we compare the keywords extracted from the future work section with those in the title and abstract texts of the same paper. If any keyword is shared by these two sections, we hypothesize that the paper and its future work section have a closer conceptual relationship. We also analyze the aggregation of this pattern on the domain level, which has been pointed out to be an indication of the maturity of a scientific field (Badua, Previts, & Vasarhelyi, 2011; Lin, 1997) and the level of interdisciplinarity of specific research communities (Rafols & Meyer, 2007; Xu, Chau, & Tan, 2014). We calculate the percentage of papers with matched keywords between title-abstract and future work sentences, in order to evaluate how these same ideas can be applied to connections between diachronic topics related to the same paper. In the following section, we offered a detailed description of our methodological framework, after which we describe our data and present some preliminary findings as a showcase of the value of our methods. The contributions of this NLP pipeline are discussed at the end of the paper.

Methods

In order to understand future work sections in scientific papers and their conceptual connections to the papers, we developed a text-analysis pipeline to identify future work statements from full-text scientific publications and extract keywords from such statements.
Part of the pipeline is adapted from the one used in our previous work (Li & Yan, in press). Each step of the pipeline is described in more detail in this section.

Within the scope of this study, we operationalize a future work statement as any description in the scientific paper specifying potentially interesting and important research topics for further investigation. We acknowledge that this definition cannot always exclusively distinguish a future work statement from other scientific paper sections bearing similar functions, especially a statement of limitations. However, any statement containing a perceived future research topic can be used to help us better understand how future work and the present research topics are framed within the same scientific paper.

Step 1: Future-work section identification: Our pipeline applies to the full-text of individual scientific papers. In order to increase the accuracy of our algorithm, we restricted our analysis to those paper sections where researchers are most likely to mention future works. We selected Discussion, Conclusion, and any other section whose title has the word “future” in it. During our preliminary work, we did find researchers occasionally mentioning future directions in the Results section or even the Introduction. We did not, however, analyze these sections, because the small number of relevant statements did not justify the potential introduction of noise into our results.

Step 2: Future-work sentence identification: In this step, we identified future work statements on the sentence level, using an algorithm to match keywords, language patterns around keywords, and syntactic features of the sentences.

Prior to the development of the algorithm, a coder reviewed 100 randomly-sampled papers from *Science Advances* to identify the most relevant terminological and syntactical patterns in future work statements. Based on this review, 12 words ("future," "further," "additional," "will," "would," "additional," "need," "needs," "needed," "remain," "warrant," and "warrants") were selected as seed terms: sentences with these terms will be pulled into the NLP pipeline for further evaluations.

For each target term in the selected sentences, rules were developed to identify the existence (or nonexistence) of other important terms next to it. A detailed list of the rules is included in the *Rule_Description* file as part of our algorithm deposited on GitHub (Li, 2018). We also excluded any sentence in which past or perfect tenses predominated. Any sentence meeting the threshold for approval was regarded as a true future work statement.

Nonetheless, the practice of selecting the whole sentence as a future work statement had two notable shortcomings. The first is that even though some sentences are true pointers of future research directions, they fail to identify any specific research topic. The following sentence is an example:

> *This issue needs to be addressed in future research.*

In this case, the true research topic we want to analyze are mentioned in a sentence prior to the one identified by the algorithm. To accommodate this situation, we decided to include an extra sentence prior to the target sentence whenever the target sentence was shorter than 16 terms and contained "this" or "these."

The other situation is that a single sentence sometimes contains both future work statements and present findings. For example:

> *Although more work is needed to determine the absolute duration of a single dose of ASNase, we demonstrate its effectiveness in curing acute lymphoblastic leukemia.*

In this case, even though the identified sentence is indeed a description of future research, it also contains descriptions of the paper per se, which could affect the identification of keywords in the next step. However, this issue did not arise frequently, and we judged that an extra algorithm to identify individual clauses about future works would introduce a great deal of unnecessary noise into the dataset. Consequently, we did not further refine the pipeline to address this situation.

After the algorithm was developed, it was tested on another random sample of 100 papers and compared with the results from manual annotation by the coder. The obtained precision is 96.4%, and recall is 76.8%. We deem this result to be satisfying, given the fact that we judge high precision to be more important than high recall for the purpose of this project.

Step 3: Keyword identification: We used the Rapid Automatic Keyword Extraction (RAKE) algorithm (Rose, Engel, Cramer, & Cowley, 2010), as implemented in the "slowraker" R package (Baker, N.P.), to extract keywords from all identified future work sentences. To compare future-work-related keywords, we also extracted keywords from the titles and abstracts of the same scientific papers.

RAKE is an unsupervised and domain-independent keyword extraction algorithm. It calculates the *ratio of degree to frequency (deg(w)/freq(w))* of all keyword candidates (terms that are not in the stop word list) in the sentence, by taking into consideration the term frequency and the length of phrase a term is part of. The higher the score is, the more likely a term is a keyword. As a result, this algorithm favors terms that are more likely to re-occur in longer phrases.

Moreover, this algorithm adds up all scores of terms in an *n*-gram; longer phrases without a stop word are normally assigned higher scores (Rose et al., 2010).

We combined titles, abstracts, and future work statements from the same paper as the input for the RAKE algorithm. This decision stems from the fact that the frequency of a term is an important variable in determining its final score for RAKE. Using a combined text from both title-abstract and future work texts stands to make the scoring on both sides more mutually consistent.

Our list of stop words consisted of the following three components:

A. Any terms that are not nouns,

B. A list of frequently used noun terms connected to future work (such as "study," "work," and "research") based on heuristics, and

C. A list of frequently used stop words in English first developed and adopted in the SMART information system at Cornell University (Buckley, Salton, & Allan, 1993) and later adopted in the "*tm"* R package (Feinerer & Hornik, 2017).

Justeson & Katz (1995) reported that adjectives, nouns, and prepositions were the chief components of keywords they examined. Despite this finding, we limited part-of-speech (POS) tags to nouns alone, because it could reduce the noises by introducing adjectives and prepositions to the results. Allowing adjectives, for instance, led RAKE to identify many phrases which were too general to be keywords, such as "wide range." Following the suggestions of Justeson and Katz, we only accepted phrases containing 2 to 4 terms as keywords. This decision mitigates RAKE's tendency to give higher scores to longer phrases, when phrases longer than 4 terms are in fact less likely to be keywords.

After the algorithm was applied to the selected texts, we conducted a baseline test to evaluate the quality of extracted keywords. A total of 244 keywords extracted from the future work statement in a sample of 54 randomly-selected articles were manually reviewed by two coders. Three scores were assigned to the keywords based on the following criteria:

- **Level 1 (score 1):** the extracted phrase is meaningful and is able to reflect the topic of the original text. Such examples include phrases representing the concepts discussed in the abstract.

- **Level 2 (score 0.5):** the extracted phrase is meaningful but cannot reflect the topic of the original text. Most of these examples are keyword-like, but too general compared to the concepts in the abstract, such as "relevant mechanism" and "theoretical investigation."

- **Level 3 (score 0):** the extracted phrase is meaningless. Most of these phrases can be easily distinguished from keywords. They were extracted because POS attributes of some words were mistakenly identified by the algorithm. Examples of this type of phrases include "present work" and "current study."

Given the ordinal levels in our rating system, we tested inter-coder reliability using Cohen's weighted kappa (Ben-David, 2008), instead of the regular Cohen's kappa method. Applying this method, the resulting Cohen's kappa (κ) is 0.724. Based on the guidelines from Landis & Koch (1977), a kappa (κ) of .724 represents a substantial strength of agreement.

Based on the results of manual coding, the mean value of keyword meaningfulness across the 54 future work statement is 0.767. Among all 244 extracted phrases, 176 (72.1%) of them were rated into Level 1. These results suggest that most of the automatically extracted keyword are

meaningful enough to represent the concepts identified as future research directions by paper authors.

Step 4: Data analysis: Once keywords were identified, we classified keywords based on their sources and compared keywords from title-abstract and future work statement. Any identical phrase used in both sources of the same paper makes the paper a matched paper, as compared to those with no identical phrase. This binary attribute of the paper was used in the analysis as a representation of the conceptual closeness between title-abstract and future work statement of the same paper.

All data collection, processing, and analysis steps described in this section were undertaken using the statistical software R (R Core Team, 2016), along with the specific R packages mentioned throughout the paper. Our algorithm is freely available on GitHub under the MIT License (Li, 2018).

Materials

For a pilot test of our algorithm, we selected *Science Advances*, published by the American Association for the Advancement of Science (AAAS). This journal was selected based on the following considerations. First, *Science Advances* is an open-access journal, whose contents are freely available at the PubMed Central (PMC) database; this satisfies the need to examine the full text of scientific papers. Second, papers published in *Science Advances* cover a broad array of scientific fields, facilitating a cross-domain comparison of future work statements. Third, papers in this journal are unambiguously classified by scientific field: each paper has a single subject term. Based on these criteria, we selected *Science Advances* over other data sources, such as *Proceedings of the National Academy of Sciences* (full texts not available to be downloaded in bulks), *Nature Communications* (no discipline classification metadata), the *Public Library of Science (PLoS)* (publications strongly skewed toward biomedicine), and the full corpus of the PubMed Central Open Access Subset (paper-level classification unavailable).

We searched and downloaded the full-text data and metadata (in XML format) of all papers in *Science Advances* via the PMC user interface. In total, 1,672 items were found. *Science Advances* began publication in 2015[2], but three papers were reported as published in 2014; these were removed from our sample. We also excluded all papers that are not research articles, leaving a final sample size of 1,620 papers. Our data sample was collected on March 22, 2018. Table 1 presents the publication year of all the sampled papers.

Table 1: Distribution of sample by publication year

Publication year	Count
2015	319
2016	480
2017	734
2018	87

To understand whether and how the conceptual connections between papers and their future work sections vary by knowledge domains, we used the over-line keywords (i.e., disciplinary keywords assigned by the journal) of *Science Advances* papers available on PMC as the

[2] http://advances.sciencemag.org/content/by/year/advances%3B2015

representation of their scientific fields. We further mapped these keywords to an in-house classification scheme inspired by the Science, Technology, Engineering, and Mathematics (STEM) classification system. Table 2 lists the seven categories in our scheme and the number of papers classified into each knowledge domain. The largest four categories (Science, Engineering, Ecological and Agricultural Science, and Medical Science), containing 1,512 articles in total, formed the subject of our domain analysis.

Table 2: Distribution of sample by publication year

Knowledge domain	Count
Science	551
Engineering	408
Ecological and Agricultural Science	380
Medical Science	173
Technology	48
Social Science	47
Mathematics	13

Distribution of analysed sections and identified keywords

In total, 537 papers from all 1,620 papers (31.4%) were found to have future work statements. Among all of the sample, 218 papers (13.5%) have any matched keyword in the future work sentences with their titles or abstracts.

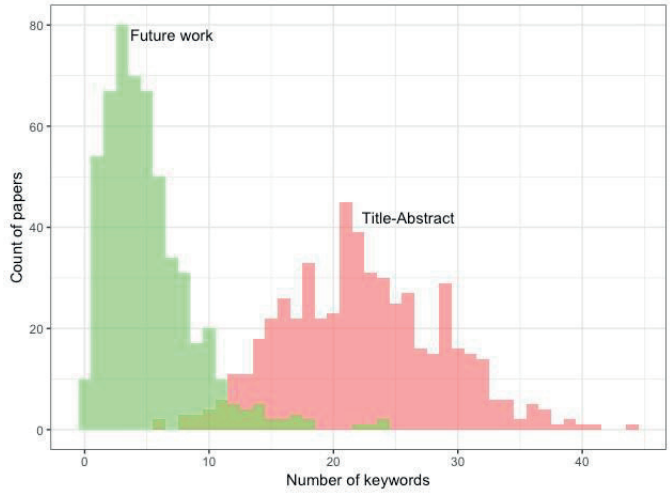
It is expected that the level of keyword matchedness is a function of the number of extracted keywords from both sources. In terms of the future work keywords, 1,241 *instances* of 1,174 unique keywords were identified from future work sentences. This ratio of unique keywords (95%) is comparable to that of the title-abstract keyword list where 10,896 unique keywords were used 11,978 times (91%). In both cases, there is a very long tail in the end of the keyword distributions, due to the small sample size. All future work keywords that were used over two times were listed in Table 3.

Table 3: The most frequently reoccurring keywords in the future work section

Keyword	Count
Climate change	7
Gene expression	5
Life history	4
Device structure	3
Mouse model	3
Protein stability	3
Size effect	3
Time scale	3

The distribution of both kinds of keywords on the paper level is summarized in Figure 1. Both following a normal distribution, their difference in size can be attributed to the variant lengths of future work and title-abstract texts. It should also be noted that there is a strong correlation between the length of texts and keywords in both future work and title-abstract: the two values are 0.863 and 0.654, respectively, and both being statistically significant.

Figure 1: Distributions of future work and title-abstract keywords.



Moreover, Table 4 and 5 also offer the summary of data by knowledge domain and publication year. It can be observed that papers have relatively similar numbers of extracted keywords across both dimensions.

Table 4: Summary of data by knowledge domain (Eco: Ecological and Agricultural Science; Eng: Engineering; Med: Medical Science; Sci: Science)

Domain	Papers	Mean titleabstract terms	Mean future work terms	Ratio of papers with future work	Ratio of matched papers
Eco	380	21.9	5.1	34.5%	13.4%
Eng	408	23.1	4.4	26%	8.6%
Med	173	21.8	6.1	50.3%	28.9%
Sci	551	22.2	4.8	30.6%	11.8%

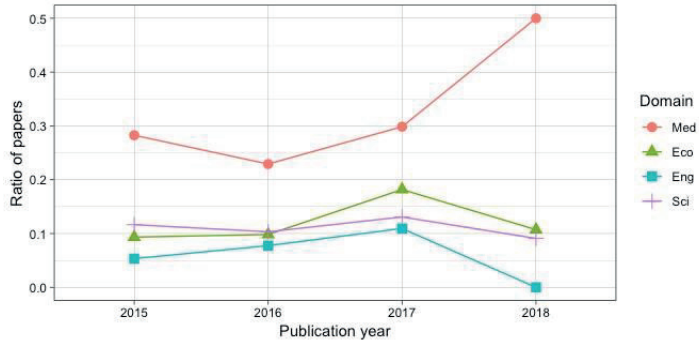
Table 5: Summary of data by publication year

Publication year	Papers	Mean titleabstract terms	Mean future work terms	Ratio of papers with future work	Ratio of matched papers
2015	299	22.1	5.4	31.1%	13.3%
2016	444	22.1	4.7	33.3%	13.5%
2017	684	22.4	5.1	39.3%	16.7%
2018	85	22.8	5.7	31.8%	16.5%

Results

The percentage of papers with matched keywords varies considerably by domain, as well as diachronically within domains. The overall percentages for our four largest domains are (from high to low): Medical Science: 28.9%; Ecological and Agricultural Science: 13.4%; Science: 11.8%; Engineering: 8.6%. Figure 1 presents the diachronic view, with a separate line for each domain. Based on the graph, there seems to be an overall upward trend in the ratio of matched papers among the four top domains between 2015 and 2017. The changes in 2018, however, may be affected by the smaller number of papers published so far this year.

Figure 2: Change over time in ratio of papers with identified future statement keywords in each domain



We conducted a series of logistic regression tests to evaluate the extent to which domain and publication year could affect the likelihood of keyword matching on the paper level. When we used knowledge domain as the only independent variable, as shown in Table 6, there is a significant difference in the likelihood of the presence of matched keywords across domains. Specifically, using Ecological and Agricultural Science as the baseline, papers in Engineering are significantly less likely, or 0.95 times as likely (95% CI [0.91, 0.998]), to have matched keywords. Papers in Medical Science are significantly more likely, or 1.16 times (95% CI [1.10, 1.24]) as likely. We found no significant difference between the categories of Ecological and Agricultural Science and Science.

Table 6: Results of Logistic Regression Model

	B	S.E.	Wald	df	p-value	O.R.	2.5%	97.5%
(Intercept)	0.134	0.017	8.813	1	1.03e-14	1.144	1.106	1.183
Class-Eng	-0.048	0.023	-2.676	3	0.043	0.952	0.909	0.998
Class-Med	0.155	0.031	0.031	3	5.18e-07	1.167	1.100	1.240
Class-Sci	-0.016	0.023	0.022	3	0.467	0.984	0.942	1.028

Publication year, by itself, is not a reliable predictor of whether a keyword in the future statement can be matched with the topic of the paper per se, according to the results shown in Table 7.

Table 7: Results of Logistic Regression Model

	B	S.E.	Wald	df	p-value	O.R.	2.5%	97.5%
(Intercept)	-32.492	20.456	-1.686	1	0.092	1.05e-15	4.06e-33	270.76
Pub.year	0.017	0.01	1.693	1	0.091	1.017	0.997	1.038

However, the story is different when both knowledge domain and publication year are considered as independent variables. As shown in Table 8, reintroducing knowledge domain back to the model reveals a clear trend in which later articles are 1.02 times (95% CI [1.00, 1.04]) as likely to have matched keywords as are articles published in the previous year. This analysis does not change our previous conclusions about knowledge domains (cf. Table 6).

Table 8: Results of Logistic Regression Model

	B	S.E.	Wald	df	p-value	O.R.	2.5%	97.5%
(Intercept)	-39.602	20.187	-1.962	1	0.050	6.32e-18	4.15e-35	0.964
Class-Eng	-0.048	0.024	-1.999	3	0.046	0.953	0.910	0.999
Class-Med	0.157	0.031	5.122	3	3.42e-07	1.170	1.102	1.243
Class-Sci	-0.017	0.022	-0.746	3	0.456	0.983	0.941	1.027
Pub.year	0.020	0.010	1.968	1	0.049	1.020	1.000	1.040

Overall, the statistical results point to significant disciplinary differences in term of the percentage of papers with matched keywords: Medical Science papers have the highest match rate, while Engineering papers have the lowest. When both publication time and disciplinary context are considered, papers published later are more likely to have matched keywords.

Discussion and conclusion

This paper presented an NLP pipeline to identify the future work sections of scientific papers and extract keywords from those sections. This represents, to our knowledge, the first empirical effort to directly investigate this important yet long-overlooked subgenre of academic writing. The technical specifications of this pipeline were detailed in this paper, and a pilot study, based on a sample of papers published in *Science Advances*, was conducted to evaluate and showcase the pipeline's usefulness.

The most important contribution of the present study is the introduction of future work statements as a research object in the field of quantitative studies of science. In this capacity, future work statements are a potentially invaluable complement to the existing data sources used to predict the future of scientific studies. Despite some successes achieved by scientometricians aiming to predict emerging topics (Haslam et al., 2008; Small, 2006), the majority of studies in this area continue to be based on citation data. Yet, as shown by various studies on the context of citation linkage, a citation or co-citation relationship does not necessarily suggest that papers share similar conceptual attributes (Cronin, 1984; MacRoberts & MacRoberts, 1989; Zhang, Ding, & Milojević, 2013). More importantly, citation data is always retrospective in nature: it is a record of an author's past decision concerning research objects created even earlier. As a result, these data give scholars limited insight into the future popularity of research topics. Future work statements, in contrast, are direct descriptions of research topics which the authors deem important to be pursued in the future. Directly analyzing authors' descriptions of future directions could greatly contribute to existing efforts to investigate future research topics. Similarly, such an analysis might inform the science policymaking and planning which are based on these scientometric studies.

Much work remains before future work sections are truly incorporated into the knowledge framework with which we analyze scientific texts. One important question to be answered is: to what extent do future work statements reflect true shifts in the conceptual landscape of scientific studies? Our pilot study offers some initial evidence which may be helpful in answering this question. We demonstrated that there are variations across knowledge domain boundaries in terms of the degree to which concepts are shared by the title/abstract and future work section of the same paper. The percentage of papers in *Medical Science* is the highest among the four domains, with *Engineering* the lowest. Moreover, the level of keyword matchedness is also shifting over time.

In addition to the two factors identified in the literature as affecting the diversity of research topics (i.e., maturity and interdisciplinarity of a field), another factor seems at least equally important in the case of future work keywords: namely, the degree of theory-driven orientation of a knowledge domain. As shown in our results, the practical nature of engineering research makes its publications less likely to have future work statement and weaker conceptual connections between future work and paper abstract. Medical sciences, on the other hand, were built on a much stronger relationship between topics pursued in publications and those perceived to be important to be pursued in the future. Such conclusions, however, needs more empirical evidence to be approved.

We hope our methodological framework will be applied in future predictive studies of science. Doing so stands to enrich our knowledge about the conceptual connections between works of different times and facilitate the practical effort of science policymaking. At the same time, the results supplied by our method must be contextualized and reevaluated with the help of other quantitative and qualitative methods, so that we can gain deeper insights into the role and nature of future work sections.

Acknowledgments

This project was made possible in part by the Institute of Museum and Library Services (Grant Award Number: RE-07-15-0060-15), for the project titled "Building an entity-based research framework to enhance digital services on knowledge discovery and delivery".

References

Badua, F. A., Previts, G. J., & Vasarhelyi, M. A. (2011). Tracing the development of accounting thought by analyzing content, communication, and quality in accounting research over time. *Accounting Historians Journal*, *38*(1), 31–56.

Ben-David, A. (2008). Comparison of classification accuracy using Cohen's Weighted Kappa. *Expert Systems with Applications*, *34*(2), 825–832. https://doi.org/10.1016/j.eswa.2006.10.022

Buckley, C., Salton, G., & Allan, J. (1993). Automatic retrieval with locality information using SMART. In *Proceedings of the First Text REtrieval Conference TREC-1* (pp. 59–72).

Cronin, B. (1984). The citation process. {The} role and significance of citations in scientific communication. *London: Taylor Graham, 1984*, *1*. Retrieved from http://adsabs.harvard.edu/abs/1984cprs.book.....C

Cross, C., & Oppenheim, C. (2006). A genre analysis of scientific abstracts. *Journal of Documentation*, *62*(4), 428–446.

Dos Santos, M. B. (1996). The textual organization of research paper abstracts in applied linguistics. *Text-Interdisciplinary Journal for the Study of Discourse*, *16*(4), 481–500.

Feinerer, I., & Hornik, K. (2017). tm: Text mining package. *R Package Version 0.7-3*.

Gledhill, C. (2000). The discourse function of collocation in research article introductions. *English for Specific Purposes*, *19*(2), 115–135.

Haslam, N., Ban, L., Kaufmann, L., Loughnan, S., Peters, K., Whelan, J., & Wilson, S. (2008). What makes an article influential? Predicting impact in social and personality psychology. *Scientometrics*, *76*(1), 169–185.

Huth, E. J. (1987). Structured abstracts for papers reporting clinical trials. *Annals of Internal Medicine*, *106*(4), 626–627. Retrieved from http://annals.org/article.aspx?articleid=701794

Justeson, J. S., & Katz, S. M. (1995). Technical terminology: some linguistic properties and an algorithm for identification in text. *Natural Language Engineering*, *1*(1), 9–27. https://doi.org/10.1017/S1351324900000048

Ke, Q., Ferrara, E., Radicchi, F., & Flammini, A. (2015). Defining and identifying sleeping beauties in science. *Proceedings of the National Academy of Sciences*, 201424329.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*(1), 159–174. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/843571

Li, K. (2018, April). nalsi/keyword_extraction: {Keyword} extraction from full-text scientific papers. https://doi.org/DOI: 10.5281/zenodo.2528641

Li, K., & Yan, E. (In press). Are NIH-funded publications fulfilling the proposed research? An examination of concept-matchedness between NIH research grants and their supported publications. *Journal of Informetrics*.

Lin, Y. (1997). *The intellectual structure of political communication research: {An} author co-citation analysis.*

MacRoberts, M. H., & MacRoberts, B. R. (1989). Problems of citation analysis: A critical review. *Journal of the American Society for Information Science*, *40*(5), 342.

Pho, P. D. (2008). Research article abstracts in applied linguistics and educational technology: {A} study of linguistic realizations of rhetorical structure and authorial stance. *Discourse Studies*, *10*(2), 231–250.

R Core Team. (2016). R: {A} {Language} and {Environment} for {Statistical} {Computing}. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.r-project.org/

Rafols, I., & Meyer, M. (2007). Diversity measures and network centralities as indicators of interdisciplinarity: case studies in bionanoscience. In *Proceedings of {ISSI}* (Vol. 2, pp. 631–637).

Rose, S., Engel, D., Cramer, N., & Cowley, W. (2010). Automatic Keyword Extraction from Individual Documents. In M. W. Berry & J. Kogan (Eds.), *Text Mining* (pp. 1–20). John Wiley & Sons, Ltd.

Samraj, B. (2002). Introductions in research articles: {Variations} across disciplines. *English for Specific Purposes*, *21*(1), 1–17.

Samraj, B. (2005). An exploration of a genre set: {Research} article abstracts and introductions in two disciplines. *English for Specific Purposes*, *24*(2), 141–156.

Small, H. (2006). Tracking and predicting growth areas in science. *Scientometrics*, *68*(3), 595–610.

Sollaci, L. B., & Pereira, M. G. (2004). The introduction, methods, results, and discussion ({IMRAD}) structure: a fifty-year survey. *Journal of the Medical Library Association*, *92*(3), 364.

Teufel, S. (2017). Do "{Future} {Work}" sections have a purpose? {Citation} links and entailment for global scientometric questions.

van Raan, A. F. J. (2004). Sleeping Beauties in science. *Scientometrics*, *59*(3), 467–472. https://doi.org/10.1023/B:SCIE.0000018543.82441.f1

Xu, J., Chau, M., & Tan, B. C. Y. (2014). The development of social capital in the collaboration network of information systems scholars. *Journal of the Association for Information Systems*, *15*(12), 835.

Zhang, G., Ding, Y., & Milojević, S. (2013). Citation content analysis (cca): {A} framework for syntactic and semantic analysis of citation content. *Journal of the American Society for Information Science and Technology*, *64*(7), 1490–1503. Retrieved from http://onlinelibrary.wiley.com/doi/10.1002/asi.22850/full

The ASEAN University Network Research Performance: A Meso-level Scientometric Assessment

Mohammadamin Erfanmanesh[1], Niusha Zohoorian-Fooladi[2] and A.Abrizah[3]

[1] *amin.erfanmanesh@gmail.com*
Department of Library and Information Science, Faculty of Computer Science and Information Technology, University of Malaya (Malaysia)

[2] *niushazohourian@gmail*
Department of Information and Library Technologies, John Abbott College (Canada)

[3] *abrizah@um.edu.my*
Department of Library and Information Science, Faculty of Computer Science and Information Technology, University of Malaya (Malaysia)
Malaysian Citation Centre, Ministry of Higher Education, Putrajaya (Malaysia)

Abstract
The ASEAN University Network (AUN) is a South East Asian (SEA) association of institutions founded in 1995 and currently has 30 members from 13 countries. This study investigates research performance of the AUN member universities over a five-year time span (2013-2017) and benchmarks their research productivity and impact, research excellence, innovation, collaboration and prominence, employing multidimensional indicators and established bibliometric methods. Data were obtained from Elsevier's Scopus and analysed utilizing SciVal. Findings show that two leading Singaporean universities accounted for almost one third of all AUN members' outputs, followed by four Malaysian universities. Singaporean universities also garnered the highest amount of citations, mean citations per publication (CPP), citedness rate, field-weighted citation impact (FWCI) and H5-index. Only 12 universities had relative citation impacts above the expected global average. A total of 999 scientific outputs of the AUN members have been cited 1674 times by the 1550 patents issued by five international patent offices. Most of the AUN members' outputs were authored jointly by two or more authors (92.48%), while single-authored publications only constituted about 7.52% of all publications. The results revealed that the AUN members published the highest share of collaborative publications in partnership with the researchers affiliated with the institutions from Asia Pacific, Europe and North America. Research prominence of each of the AUN members were studied and visualized using direct citation analysis of their scientific output.

Introduction
Research within Asia and internationally is propelling Asian countries to new economic heights and trends in research output are already showing a rise in Asia's share. There has been a shift from national to bilateral research projects and, more recently, a trend that sees a number of countries in the Asian region engaging in research collaboration. As a result, top Asian countries such as China, Japan, Hong Kong, Singapore, and South Korea have made significant gains in research performance (Erfanmanesh et al., 2013). However, the same is not true in most Southeast Asian (SEA) countries and universities (Barrot, 2017). Based on SCImago Journal and Country Rank (SJR) data, the research productivity of SEA countries in all disciplines (827,809) only comprise 1.69 percent of the worldwide productivity (48,969,648) and only 1.37 percent in terms of citations excluding self-citations[1]. Most universities in the SEA region (except Singapore and Malaysia) have ranked poorly as well in several university rankings (e.g. Times Higher Education and QS) that heavily consider research productivity and impact. Nevertheless, an analysis mapping of research trends and collaboration in the Association of South East Asian Nations (ASEAN) region showed that ASEAN's share of the world literature has been almost doubled from 1.37% in 2006 to 2.43% in 2015 (Thomson Reuters, 2016) with a large proportion are resulted from international collaboration.

The 4th ASEAN Summit in 1992 called for ASEAN Member Countries to help "hasten the solidarity and development of a regional identity through the promotion of human resource development so as to further strengthen the existing network of leading universities and institutions of higher learning in the region".[2] This idea led to the establishment of the ASEAN University Network (AUN) in November 1995 with the signing of its Charter by the Ministers responsible for higher education from six Member Countries, and with initial participation of eleven universities from six countries. AUN's strategic focus related to research built on those identified by ASEAN are twofold: (a) To promote collaborative study, research and educational programmes in the priority areas identified by ASEAN; (b) To promote cooperation and solidarity among scholars, academicians and researchers in the ASEAN Member States. Currently AUN consists of 30 universities from 13 countries; however, there is a great variation of research output across members of AUN in terms of the amount of published peer-reviewed literature in the core body of scientific literature.

Some research efforts empirically studied science, technology and innovation (STI) in ASEAN countries (e.g. Dodgson, 2000; Lai and Yap, 2004; Remøe, 2010; Sigurdson and Palonka, 2002; Rodriguez and Soeparwata, 2012). Although there have been few studies evaluating the research performance of ASEAN in various fields (Sombatsompop et al., 2011; Kumar, Rohani & Ratnavelku, 2014; Payumo & Sutton, 2015) and countries (e.g. Rana (2012) and Prathap (2018) on Singapore, Kumar and Jan (2013) and Tahira et al. (2016, 2018) on Malaysia, Maula, Fuad and Utarini (2018) on Indonesia, Sombatsompop et al. (2006) on Thailand and Vinluan (2012) and Navarrete and Asio (2014) on Philippines), it appears there has been no empirical analysis of all ASEAN members at the meso-level that engages in research performance assessment. The review of previous research revealed a gap in the literature regarding research activities of the top ASEAN research institution and benchmarking their performance based on various metrics of research productivity, impact, excellence and collaboration. Therefore, this study focuses on identifying the research profiles of ASEAN countries by characterizing the trends in publication activity and citation impact at the meso-level. Specifically, it gives an overview of the relative size of scientific performance of AUN members in the 5-year period (2013-2017) and attempts to address the following research questions:

a) What is the publication productivity of AUN members in the period 2013-2017 and the scientific impact received by these publications?
b) How do AUN members compare among their peer institutions in terms of research excellence, innovation performance and research prominence?
c) What is the authorship pattern and how does this translate into collaboration network among ASEAN member countries?

Materials and Methods

The study aims to characterize the trends in publication activity among the AUN members and benchmarking research impact and excellence of these institutions over a five-year time span. The specific objectives are to examine the AUN member universities: (a) research productivity and impact; (b) research excellence; (c) innovation performance; (d) authorship patterns and research collaborations; and (e) research prominence. The current research is conducted using bibliometric approach based on the data collected from the Elsevier's Scopus in September 2018. SciVal which is the Elsevier's subscription-based analytical tool based on Scopus data was utilized for benchmarking the AUN member's research performance. Research sample comprised all Scopus-indexed publications of the AUN members during a five-year time period from 2013 through the end of 2017. The list of ASEAN University Network (AUN) member universities was extracted from its website[3]. Each university was searched in the Scopus and SciVal and its data were extracted in a Microsoft Excel spread sheet for further investigations.

Although AUN has 30 members, two institutions from Cambodia (Royal University of Law & Economics) and Myanmar, (Yangon Institute of Economics), did not have any publication indexed in the Scopus during the five-year period, and therefore were excluded from the study. Therefore, the metrics listed in Table 1 were investigated for 28 AUN member universities. Data analysis was performed using Microsoft Excel. Research prominence circle graphs were plotted using the visualization tools of the SciVal. Moreover, to visualize the collaboration network of the AUN member universities, the UCINET version 6 was utilized.

Table 1: Metrics Used in this Study

Scholarly Output:	The total number of publications of an institution in Scopus-indexed journals.
Citation Count	The total number of citations received by an institution's publications in Scopus.
Citations per Publication (CPP):	The average citation impact of an institution's publications.
Citedness Rate	The proportion of an institution's publications with at-least one citation
Field Weighted Citation Impact (FWCI):	The ratio of citations received by an institution's publications relative to the expected world average for the subject category, publication type and publication year. A FWCI of more than 1 indicates that an institution's publications have been cited more than would be expected based on the global average for the field and vice versa.
H5-Index:	The h-index of an institution based upon data from the five years under investigation in this research (2013-2017).
Outputs in the Top Citation Percentiles:	The share of an institution's scholarly output that are within the top 1%, 5%, 10% and 25% of the most cited publications in the subject categories assigned to the parent journal.
Outputs in the Top Journal Percentiles:	The share of an institution's scholarly output that are published in the top 1%, 5%, 10% and 25% journals with the highest SJR in the subject categories assigned.
Citing-patents Count:	The number of patents citing the scholarly output published by an institution.
Patent-cited Scholarly Output:	The number of an institution's scholarly output that have been cited in patents.
Patent-citations Count:	The total number of patent citations received by an institution's scholarly output (SciVal Research Metrics Guidebook, 2018).

Findings

Research Productivity and Impact

The first of the performance indicators considered here is research productivity and impact, which can be measured by the use the following indicators: publication count, citation count, CPP, FWCI, citedness rate and H5-index (Table 2). Results showed that two universities from Singapore (NUS and NTU), had the highest research productivity during the five years under study. Four Malaysian universities (UM, UPM, UKM and USM) ranked third to sixth. The total number of unique authors affiliated with each institution with at least one publication was also depicted in Table 2. Considering the mean publications per researcher, SMU had the best performance with the mean value of 3 publications for each author, followed by the NTU (2.26) and UM (1.88). It is apparent that the most number of citations received go to the institutions with the highest paper count (NUS, NTU and UM). These three universities also recorded the top most CPP within the analysed time frame. Over 66% of all papers published by the AUN members had been cited at least once. The highest citedness rates were seen in NUS, NTU and RUPP. In terms of FWCI, NTU (1.89) had a higher value than the other universities, while UUM (0.49) showed an extremely low value. A FWCI of 1.89 for the NTU indicates that this institution's publications have been cited 89% more than the expected citation impact based on the global average. Only 12 universities had relative citation impacts above the expected global average. Consideration was also given to the H5-index. NTU (170) and NUS (164) were observed to have the highest H5-index. The lowest value belonged to MU with a H5-index of

4, which indicates that this institution has published 4 papers during the last five years and each has been cited at least 4 times.

Table 2: Research Productivity and Impact of the AUN Members

University	*Scholarly Output*	*Authors*	*Citation Count*	*CPP*	*Citedness Rate*	*FWCI*	*H5-index*
NUS	46169	24355	583505	12.6	80.7	1.76	164
NTU	35153	15588	471842	13.4	79.1	1.89	170
UM	22275	11846	201057	9	77.2	1.33	95
UPM	16479	11464	87123	5.3	68.3	0.84	60
UKM	16270	10244	81190	5	62.6	0.88	61
USM	15495	10271	84158	5.4	65.8	0.93	62
MaU	11458	9298	90485	7.9	74.8	1.18	82
CU	10830	7754	75922	7	71.4	1.06	71
CMU	6609	4364	34304	5.2	69.8	0.87	48
ITB	5794	4814	15319	2.6	52.9	0.84	30
UI	5464	5659	16402	3	50.5	0.86	35
PSU	4618	3304	26796	5.8	71.9	0.93	39
UP	4510	4712	33850	7.5	56.4	1.62	52
UGM	3799	3692	9966	2.6	56.4	0.77	23
UUM	3544	2283	6029	1.7	43.5	0.49	20
VNU-HCM	2787	2697	12739	4.6	66.2	0.91	30
VNU-HN	2325	1945	12533	5.4	72.2	1.02	30
SMU	2271	756	15908	7	74.4	1.77	37
DLSU	1620	1305	7197	4.4	58.3	1.10	27
UBD	1296	701	7650	5.9	70.1	1.42	27
UNAIR	1025	1214	2422	2.4	53	0.56	15
BUU	986	659	3625	3.7	64.2	0.66	19
CTU	657	533	3583	5.5	72.6	0.99	22
ATMU	627	533	2821	4.5	56.3	1.49	16
NUOL	191	148	984	5.2	67.5	0.99	13
RUPP	103	86	440	4.3	77.7	1.09	9
UY	71	49	359	5.1	69	0.93	9
MU	18	20	60	3.3	66.7	0.58	4

Research Excellence

To demonstrate research excellence of the AUN members, two indicators were utilized: outputs in top citation percentiles and outputs in top journal percentiles. Table 3 shows the share of each institution's publications in top 1%, 5%, 10% and 25% of the most-cited publications. Once again, NTU, NUS and UM garnered the highest share of highly-cited publications in all four thresholds. Results revealed that 4.5% of the overall papers of the NTU were among the top 1% most cited publications of their corresponding subject categories in Scopus. Three institutions (RUPP, UY and MU) had no papers amongst the world's 1% top-cited publication.

Table 3: Outputs of the AUN Members in Top Citation Percentiles

University	*Top 1% of the World (%)*	*Top 5% of the World (%)*	*Top 10% of the World (%)*	*Top 25% of the World (%)*
NTU	4.5	14.6	24.1	45.3
NUS	3.7	13.8	23.6	46.2
UM	2.4	10.1	18	38.8
UBD	2.2	6.6	12.3	30.1
MaU	1.7	6.3	12.8	34.6
ATMU	1.6	3.5	6.5	17.5
UP	1.6	4.4	8	20.5
CU	1.4	6.3	12.9	31.8
SMU	1.2	7.2	12.8	33.9
BUU	1	1.8	4.9	21.2
USM	0.8	3.9	8.5	23.7
UPM	0.7	4.2	8.9	24.9
UKM	0.7	3.7	7.3	19.9
DLSU	0.7	3.8	9.3	23
CMU	0.7	3.9	8.5	25.9
PSU	0.6	3.4	9.1	27.5
UI	0.5	2.1	4.5	14.4
NUOL	0.5	5.8	8.4	27.7

VNU-HCM	0.5	3.9	8.4	24.9
VNU-HN	0.4	4.4	11.2	30.5
ITB	0.3	1.5	3.7	13.2
CTU	0.2	4.9	10.7	30.7
UUM	0.1	0.9	2.3	7.6
UGM	0.1	1.3	4.1	15.6
UNAIR	0.1	1.2	3.1	15.5
RUPP	-	1	2.9	33
UY	-	1.4	12.7	31
MU	-	5.6	5.6	22.2

Research excellence of the AUN members based on their contribution in the world's top ranked journal was also examined. Publication output in top journal percentiles indicates the extent to which an institution's publications are present in the top 1%, 5%, 10% and 25% journals indexed by Scopus, as measured by SJR. Singapore accounts for 20.3% of the overall papers published in top 1% Scopus-indexed journals during the five-year period (Table 4), contributed by NUS (7.2%), SMU (6.8%) and NTU (6.3%).

Table 4: Outputs of the AUN Members in Top Journal Percentiles

University	*Top 1% Journals (%)*	*Top 5% Journals (%)*	*Top 10% Journals (%)*	*Top 25% Journals (%)*
NUS	7.2	26.1	45.7	72.4
SMU	6.8	22.7	33.4	57.0
NTU	6.3	23.6	41.1	67.1
RUPP	3.4	5.7	21.6	51.1
NUOL	3.1	6.7	12.9	32.5
MaU	3.0	12.2	25	54.9
UP	2.3	7.4	15.6	34.5
UBD	1.1	5.4	14.7	33.5
ATMU	1.0	6.1	11.7	25.1
CU	1.0	7.9	18.2	47.0
CMU	0.8	5.2	12.2	38.2
UI	0.7	3.6	7.5	17.9
UM	0.6	7.0	18.3	44.7
PSU	0.6	3.4	10	34.2
VNU-HCM	0.6	4.0	11.8	31.7
UGM	0.5	2.3	7.2	17.6
DLSU	0.5	3.3	9.3	21.0
UNAIR	0.4	2.4	5.2	17.0
USM	0.3	2.5	7.9	24.5
UKM	0.3	2.3	7.1	19.5
ITB	0.3	2.3	5.4	13.7
VNU-HN	0.3	5.2	15.8	44.0
BUU	0.2	3.9	8.9	31.5
CTU	0.2	3.4	15.5	40.3
UPM	0.1	2.0	6.5	23.8
UUM	0	0.9	1.7	4.0.
UY	-	3.3	6.6	42.6
MU	-	5.6	27.8	55.6

Innovation Performance

Patenting remains a powerful indicator to capture the intermediate stages of innovation activities. Three patent-related metrics were utilized to investigate academic-industry connections and knowledge flows of the AUN members, namely citing-patents count, patent-cited scholarly output and patent-citations count, as these indicators are likely to reflect the number of inventions, levels of innovation or the benefits from innovation. SciVal covers patents data from five of the largest patent offices: European Patent Office (EPO), US Patent Office (USPTO), UK IPO (UK Intellectual Property Office), JPO (Japan Patent Office) and WIPO (World Intellectual Property Organization). A total of 999 scientific outputs of the AUN members have been cited 1674 times by 1550 patents issued by the five patent offices (Table 4). With regard to all three patent-related metrics, NUS tops the list with 686 citations to 416 publications from 627 patents, followed by NTU and UM. If we consider the share of patent-

cited scholarly output to the overall publications of each institution, the ranking would be slightly different. Patent-cited publications of the NUS were responsible for 0.9% of the overall output of this institution over the five-year time period, while this proportion was found to be 0.75% for NTU and 0.4% for MaU and UM. Eight out of the 28 studied institutions did not have any patent-cited scholarly output and we can say that their research output is not being used in the creation of products as captured by the five investigated patent offices (Table 5).

Table 5: Innovation Performance of the AUN Members

University	*Citing-patent Count*	*Patent-cited Scholarly Output*	*Patent-citations Count*
NUS	627	416	686
NTU	428	263	460
UM	149	89	156
MaU	66	46	84
UPM	50	40	51
USM	45	31	47
UKM	41	28	45
CU	41	32	42
ITB	30	7	30
CMU	21	12	21
PSU	18	12	18
VNU-HCM	9	6	9
UP	8	7	8
UGM	6	3	6
SMU	4	1	4
UI	2	1	2
CTU	2	2	2
UNAIR	1	1	1
UBD	1	1	1
VNU-HN	1	1	1
UUM	-	-	-
RUPP	-	-	-
NUOL	-	-	-
UY	-	-	-
MU	-	-	-
ATMU	-	-	-
DLSU	-	-	-
BUU	-	-	-

Authorship Patterns and Research Collaboration

Joint authorship is the mirror to research collaboration among countries, and collaboration reveals knowledge flows within the knowledge system. Each publication is assigned to one of four authorship patterns (single, institutional, national and international) based on its affiliation information. Single-authorship was assigned to a paper if no research collaboration was found. International collaboration was assigned if a paper was written by authors affiliated with institutions from more than one country. Papers published with the involvement of two or more institutions from the same country were assigned as national collaboration. Moreover, papers written by more than two researchers affiliated with a single institution were considered as institutional collaboration. Table 6 shows the share of authorship and sector of collaboration patterns along with the CPP and FWCI of each authorship pattern for the AUN members. Results showed that the majority of the AUN members' publication outputs were a result of collaborative authorship (92.48%), while single-authored publications only accounted for 7.52% of all publications. Institutional, national and international collaborative publications were accounted for 31.78%, 12.81% and 47.89% of total publications, respectively. Results revealed that internationally co-authored publications of the AUN members had higher citation impact in terms of CPP and FWCI compared with other authorship patterns. Internationally collaborative papers received the mean citation rate of 8.2 which is substantially higher than that of for national collaboration (3.82), institutional collaboration (3.25) and single-authored publications (2.0). The same trend can be seen with regard to FWCI and internationally co-

authored publications had almost two-fold value compared with the nationally co-authored publications. This shows that universities with larger shares of international collaborations are, in most cases, associated with higher citation rates (Spearman $r = 0.469$, $p = 0.012$).

Table 6: Authorship Patterns of the AUN Members

University	Single Authorship			Institutional Collaboration			National Collaboration			International Collaboration		
	Share (%)	CPP	FWCI	Share (%)	CPP	FWCI	Share (%)	CPP	FWCI	Share (%)	CPP	FWCI
ATMU	28.2	0.9	0.38	29.7	1.9	0.59	6.5	1.7	0.69	35.6	10	3.26
BUU	14.1	1	0.27	21.3	1.7	0.40	36.6	4.4	0.71	28.0	5.6	0.98
CMU	4.1	1.7	0.56	36.1	3.3	0.62	21.8	4.4	0.73	38.0	7.8	1.23
CTU	6.4	1.4	0.55	13.9	1.3	0.51	8.4	3.4	0.76	71.3	6.9	1.15
CU	4	1.6	0.43	36.5	3.5	0.63	19.3	4.8	0.73	40.2	11.8	1.67
DLSU	16.4	1.5	0.40	33.1	2.1	0.77	8.5	2.6	0.53	42.0	7.8	1.75
ITB	3.8	1.4	0.69	64.4	1.5	0.79	10	1.3	0.62	21.8	6.8	1.12
MaU	3.7	2.5	0.64	30.9	3.5	0.55	19.9	4.9	0.70	45.5	12.6	1.86
MU	11.1	4.5	0.68	11.1	2.5	0.47	-	-	-	77.8	3.3	0.58
NTU	5.3	3.5	0.92	26.6	13.2	1.74	9.4	10.6	1.49	58.7	14.9	2.11
NUOL	2.1	2.5	0.93	7.3	0.6	0.20	-	-	-	90.6	5.6	1.05
NUS	7.2	3.1	0.9	21.8	9.9	1.42	13	11.4	1.46	58.0	15.2	2.06
PSU	3.9	1.9	0.41	40.4	3.8	0.63	20.5	4.9	0.78	35.2	9.1	1.40
RUPP	7.8	1.6	0.48	3.9	1.5	0.61	1	6	0.76	87.3	4.6	1.17
SMU	12.7	2.3	0.64	18.6	4.8	1.41	5.8	7.2	1.84	62.9	8.6	2.1
UBD	12.2	1.4	0.47	24.2	3.7	0.89	0.9	3.1	0.52	62.7	7.7	1.82
UGM	4.9	1.2	0.6	53.1	1.5	0.72	13.2	1.5	0.66	28.8	5.4	0.94
UI	4.9	1.3	0.48	62.4	1.4	0.64	9.3	1.3	0.54	23.4	8.4	1.64
UKM	2.6	2.3	0.53	43.3	3.4	0.59	26.1	3.6	0.64	28	9	1.61
UM	3.3	2	0.6	29.4	7.4	1.03	18.7	6.2	0.93	48.6	11.6	1.71
UNAIR	4.7	0.6	0.29	46.6	1.1	0.40	18	1.4	0.43	30.7	5.2	0.93
UP	13.4	1.1	0.44	37.2	1.4	0.39	5	2.6	0.55	44.4	15.2	3.12
UPM	1.4	4.1	0.64	36.7	4.1	0.66	24.3	4.9	0.78	37.6	6.8	1.07
USM	2.6	2.2	0.62	40.2	4	0.65	21.7	4.5	0.76	35.5	7.9	1.37
UUM	5.1	1.6	0.33	50	1.3	0.42	20.9	1.6	0.40	24	2.7	0.73
UY	9.9	0.4	0.11	9.9	-	-	-	-	-	80.2	6.2	1.15
VNU-HCM	5	2.7	0.55	39.3	2.7	0.65	7.4	5.1	1	48.3	6.2	1.15
VNU-HN	9.8	3.8	0.98	22	3.9	0.80	12.5	3.6	0.84	55.7	6.7	1.16

Table 7 shows the number of collaborating institutions and collaborative publications of the AUN members based on six geographic regions. Analysis of authors' nationalities showed notable variations in the geographic distribution of collaborations in each AUN member. Nearly 45% of AUN research has been published in collaboration with institutions from the Asia Pacific and another 13.5% and 11.5% with European and North American partners respectively. In contrast, institutions from the Middle East, Africa and South America had the lowest amount of joint publications with the AUN members. This shows that AUN members with larger shares of international collaborations are, in most cases, associated with countries that are investing heavily in research, such as China, Korea, Japan, US and the UK.

Table 8 shows AUN members' top collaborating institutions with the highest co-authorships and the top collaborating institutions with the greatest FWCI. Only 5 out of the 28 AUN members had the highest number of collaborations with institutions from different country (CTU-Taiwan; MU-Japan; NUOL-Thailand/Japan; SMU-China; UNAIR-Japan). However, for all AUN members, the highest amount of FWCI was found to belong to the papers which have been published jointly with foreign institutions. Figure 1 plots the AUN members' collaboration network during the five-year period. The co-authorship network consists of nodes and links: nodes represent institutions, while links connect nodes in terms of co-authorships. There is a link between institutions if they have co-authored at-least one publication. The size of a node is proportional to the number of co-authorships that a given institution has in the network. Moreover, the strength of links between two nodes demonstrates the amount of co-authorships of those institutions. The collaborating ties between institutions from the same countries are noticeable in Figure 1, i.e. NTU and NUS (Singapore); UM, UKM, UPM, USM (Malaysia) and CMU, MaU and CU (Thailand).

Table 7: Geographic Distribution of the AUN Members' International Collaboration

University	*North America*		*South America*		*Europe*		*Asia Pacific*		*Middle East*		*Africa*	
	Coll. Inst.	*Coll. Pub.*	*Coll. Inst.*	*Coll. Pub.*	*Coll. Inst.*	*Coll. Pub.*	*Coll. Inst.*	*Coll. Pub.*	*Coll. Inst.*	*Coll. Pub.*	*Coll. Inst.*	*Coll. Pub.*
ATMU	203	92	57	16	234	69	187	185	48	28	55	26
BUU	73	91	8	11	115	95	147	538	4	6	3	5
CMU	325	823	55	87	486	814	565	3035	59	109	53	130
CTU	28	28	2	3	105	214	169	307	3	3	3	8
CU	463	1872	81	594	768	1595	757	4851	95	546	64	359
DLSU	151	130	47	17	208	93	273	711	24	21	29	17
ITB	164	127	51	8	280	386	306	1444	24	36	18	16
MaU	520	2240	113	241	911	2194	804	5191	119	199	125	625
MU	2	2	-	-	10	4	13	11	1	1	-	-
NTU	649	5003	98	164	1237	6112	1297	17180	176	1062	93	132
NUOL	42	25	18	3	88	58	123	132	2	2	25	4
NUS	826	8871	162	428	1570	8252	1494	23495	211	1203	150	371
PSU	217	373	25	25	401	557	425	1910	41	89	34	39
RUPP	13	13	-	-	47	39	64	58	-	-	1	1
SMU	223	665	19	14	215	298	234	914	21	30	13	9
UBD	108	127	28	17	250	259	274	524	25	50	23	38
UGM	86	94	14	13	226	420	293	1244	13	25	29	31
UI	205	269	53	31	533	548	462	1320	68	58	82	52
UKM	316	383	65	66	615	940	631	6658	193	1286	103	210
UM	494	2278	133	946	1089	3335	1044	10992	270	3101	145	1150
UNAIR	27	56	3	4	73	58	190	455	8	11	7	8
UP	397	592	91	113	591	609	580	1466	95	106	95	115
UPM	266	472	65	28	526	1102	597	6845	227	2175	101	484
USM	402	570	100	53	787	962	809	6258	205	1531	128	426
UUM	38	70	8	7	90	105	137	1068	65	159	26	65
UY	1	1	-	-	8	6	54	53	1	1	-	-
VNU-HCM	119	179	10	13	294	466	311	1072	22	40	10	15
VNU-HN	133	280	31	80	358	576	300	1015	17	51	20	24

Table 8: Top Collaborating Institutions with the Highest Publications and Citation Impact

University	***Top Collaborating institutions with the highest co-authorships***	***Top Collaborating institutions with the highest FWCI***
ATMU	University of the Philippines (49)	University of New Mexico and 11 more Institutions (95.60)
BUU	Mahidol University (128)	Friedrich Schiller University Jena (17.86)
CMU	Mahidol University (428)	Landcare Research (9.34)
CTU	**National Taiwan University of Science & Technology (54)**	Universite de Toulouse (7.39)
CU	Mahidol University (744)	University of Minnesota (5.18)
DLSU	University of the Philippines (132)	G.B. Pant Hospital India (18.87)
ITB	Lembaga Ilum Pengetahuan Indonesia (150)	Harvard University (6.67)
MaU	Chulalongkorn University (744)	Medical University of Vienna (19.54)
MU	**Kyoto University (6)**	University of Oxford and four more institutions (4.35)
NTU	Agency for Science, Technology & Research Singapore (3067)	University College London (11.22)
NUOL	**Kasetsart University, Kyushu University (11)**	University of California at San Francisco and 19 more institutions (4.93)
NUS	Agency for Science, Technology & Research Singapore (5468)	University of Edinburgh (16.60)
PSU	Chiang Mai University (256)	National Taiwan University (21.33)
RUPP	CNRS, Chiang Mai University (7)	National University of Singapore (15.87)
SMU	**Zhejiang University (109)**	Macquarie University (8.70)
UBD	National University of Singapore (50)	University Technology Malaysia (33.15)
UGM	Universitas Sebelas Maret (108)	KU Leuven (3.12)
UI	Ministry of Health Indonesia (144)	University of the Witwatersrand (23.81)
UKM	University Putra Malaysia (931)	Harvard University (111.64) 23
UM	University Putra Malaysia (961)	Chinese University of Hong Kong (5.38)
UNAIR	**Kobe University (54)**	Alfred Hospital (6.32)
UP	De La Salle University (132)	Shanghai Jiao Tong University (119.65)
UPM	University of Malaya (961)	Kyushu University (16.28)
USM	University of Malaya (592)	Karolinska Institute (61.05)

UUM	Universiti Kebangsaan Malaysia (163)	University of Washington and 30 more institutions (2.92)
UY	University of Tomaya (14)	University of Leicester and 2 more institutions (7.71)
VNU-HCM	Ton Duc Thang University (166)	Ghent University (3.77)
VNU-HN	Vietnamese Academy of Science & Technology (257)	John Hopkins University (2.81)

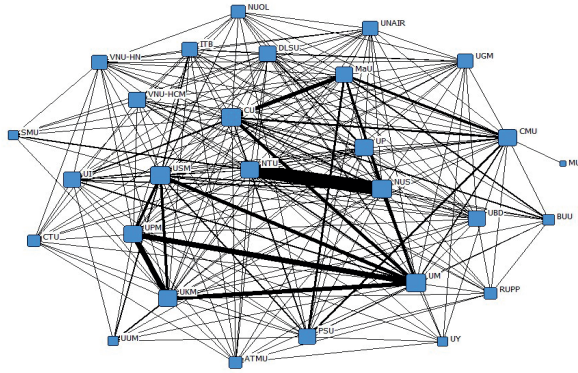


Figure 1: Collaboration Network of the AUN Member Universities

Research Prominence

Research strengths and prominence subject areas of the AUN member universities were also explored to demonstrate their scientific dominance. Subject areas with the highest publications and FWCI are presented in Table 9. It is interesting to observe a clear dominance of Engineering and Medicine, as each was the most productive subject area in 8 and 7 institutions respectively. UBD was the only AUN member university with the highest share of publications in the area of Social Sciences. It must be noted that only in 2 institutions (UP and UBD), the most productive subject areas also had the highest impact as measured by FWCI.

Table 9: Subject Categories with the Highest Publications and FWCI

University	*Most productive subject area*	*Most impact subject area (FWCI)*
ATMU	Social Sciences (263)	Medicine (14.47)
BUU	Mathematics (182)	Nursing, Veterinary (1.06)
CMU	Medicine (2005)	Environmental Science (1.67)
CTU	Agricultural & Biological Sciences (216)	Pharmacology, Toxicology & Pharmaceutics (2.01)
CU	Medicine (2885)	Physics & Astronomy (1.98)
DLSU	Engineering (472)	Immunology & Microbiology (3.13)
ITB	Engineering (2103)	Veterinary (2.47)
MaU	Medicine (6349)	Earth & Planetary Sciences (1.80)
MU	Earth & Planetary Sciences, Agricultural & Biological Sciences (6)	Earth & Planetary Sciences (1.35)
NTU	Engineering (12873)	Veterinary (2.85)
NUOL	Agricultural & Biological Sciences (53)	Medicine (2.19)
NUS	Medicine (11123)	Chemistry (2.2)
PSU	Medicine (1155)	Nursing (1.29)
RUPP	Environmental Science (36)	Arts & Humanities (8.25)
SMU	Computer Science (1088)	Multidisciplinary (2.78)
UBD	Social Sciences (285)	Social Sciences (2.67)
UGM	Engineering (989)	Dentistry (1.51)
UI	Engineering (1394)	Multidisciplinary (1.46)
UKM	Engineering (4351)	Medicine (1.75)
UM	Engineering (5298)	Decision Sciences (2.01)
UNAIR	Medicine (274)	Multidisciplinary (1.24)
UP	Medicine (1011)	Medicine (4.78)
UPM	Agricultural & Biological Sciences (3648)	Earth & Planetary Sciences (1.67)
USM	Engineering (4054)	Medicine (1.20)
UUM	Business, Management & Accounting (947)	Nursing (3.42)
UY	Chemistry (14)	Earth & Planetary Sciences (1.73)
VNU-HCM	Computer Science (930)	Arts & Humanities (3)
VNU-HN	Mathematics (637)	Health Professions (5.81)

Figure 2 illustrates much clearer pictures of research strength where an institution has a leading position compared to other institutions in terms of number of publications, citation impact, view counts and Citescore. Documents with a common intellectual interest are clustered based on direct citation analysis. These publication clusters are then grouped together into topics. Each bubble represents a topic for the studied institution. The larger the bubble size is, the more active and prominent is that institution in momentum of that topic. Nodes positioned towards the edge of the circle indicate strong distinctive topics in the color-coded subject area, while nodes positioned towards the middle indicate an interdisciplinary mix. The plots clearly show the research prominence of NTU in Engineering, NUS in Engineering and Veterinary, UM in Physics and Astronomy, UPM and USM in Materials Science and UKM in Chemical Engineering and Mathematics.

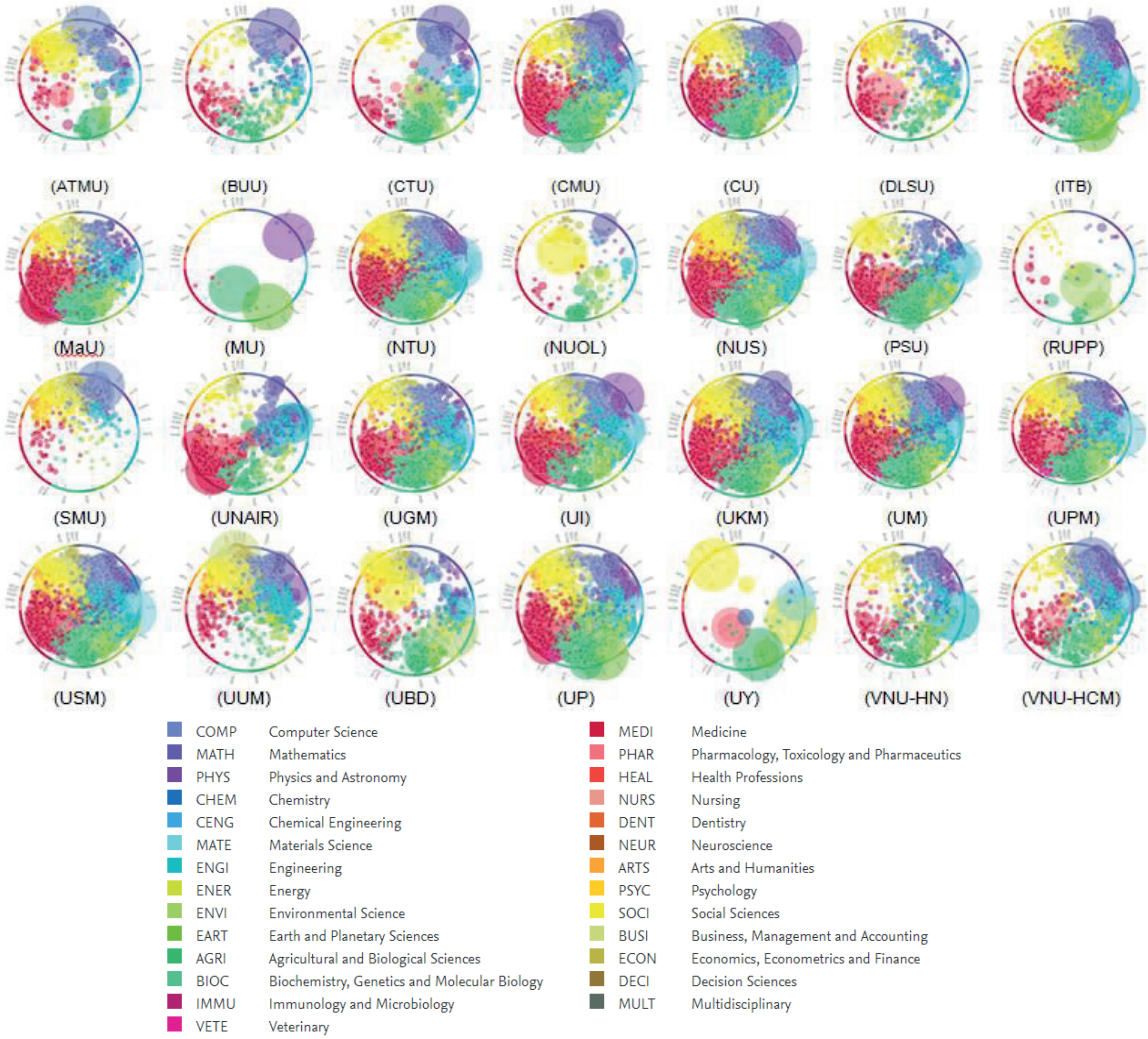


Figure 2: Topics of Research Prominence for the AUN Member Universities

Discussion and Conclusion

This study analyses the research performance of 28 AUN member universities, of which 22 have already positioned themselves in at least one global ranking systems i.e. QS World University Ranking. In general, ASEAN research base is performing strongly. Some ASEAN countries are significantly contributing to the research productivity and impact of SEA region. One possible explanation for this is that ASEAN universities, especially Malaysia, Singapore and Thailand, have begun to take very seriously the university rankings which are highly dependent on research performance. Ten AUN member universities (CMU, CU, MaU, NTU, NUS, PSU, UKM, UM, UPM and USM), as reflected from their position in ARWU, have formed a representative sample to describe and mirror the characteristics of world leading research universities from SEA. These results may be linked to the kind of faculty that these universities hire. Notably, Singapore and Malaysian universities heavily consider track record in research when recruiting academic faculty while other SEA countries remain to put prime on teaching and consider research as a second priority (Barrot, 2017). The results may also be attributed to the publication requirement for students enrolled in postgraduate degree programmes by research in these universities, where they must show proof of having published or submitted indexed-journal articles based on research conducted during their candidature prior to graduation.

As well as increasing competition, the expansion in the output of AUN member universities provides increased opportunities for international collaboration. Such collaborations allow them to leverage resources and expertise that would otherwise be unavailable. Almost half of AUN research has been published in collaboration with the Asia Pacific countries and a quarter with European and US partners. Much of this grows from a desire among AUN researchers to collaborate with their best international peers. They will go to where the opportunities are – to the US and Europe, and to countries like China, Japan, Korea and Taiwan that have invested heavily in science. As capacity and expertise build in research intensive universities worldwide are currently undergoing rapid expansion in activity, the world top universities will become increasingly selective in choosing international partners and will become, at least to some extent, more self-sufficient. Therefore, in order to remain an attractive partner, AUN member universities cannot be complacent and need to continue to support adequately its capacity to produce excellent research. This means that if ASEAN wishes to remain globally competitive it will need to maintain both research output and quality. It is likely that some universities in the AUN consortium do not normally have sufficient resources in order to be internationally active on a frequent basis, neither are there ASEAN resources that are comparable with the international programmes on cooperation and mobility. Especially for the universities that have not been positioned in the world ranking systems, it might be the first significant opportunity to cooperate regionally.

The analysis of the AUN research prominence shows that ASEAN capacity to produce excellent research seems to be concentrated in the fields of Engineering and Medical sciences and is at greater risk in some areas than in others. Research assessment exercises in ASEAN countries have resulted in resources becoming increasingly concentrated in those disciplines performing the best research, and resulted in the poorest research being cut. This has meant that in many fields ASEAN continues to produce a substantial level of increasingly high quality output. There are, however, fields where ASEAN citation impact is possibly relatively high and rising but world share is relatively low and falling. In these cases, it seems unlikely that continued improvement in citation impact can be maintained unless the fall in share is addressed. These observations, coupled with government research cuts, mean that ASEAN needs to carefully consider how best to allocate resources to maintain competitiveness and capacity in its research base. Future innovative products and services are likely to result from curiosity-driven research, the application of which is difficult to predict.

This study serves as a benchmarking tool and allows stakeholders of the AUN consortium to view and compare statistics on five dimensions of research performance, namely research productivity and impact; research excellence; innovation performance; authorship patterns and research collaboration and research prominence. It not only provides comprehensive and quantitative descriptions of AUN member universities, but it also helps these universities to position themselves at a global level from various angles and perspectives, and help them to identify their advantages and disadvantages as compared to those of their concern. As a way to extend the value of this study, further investigations are needed regarding national and institutional research policies of highly productive and impactful AUN member universities and how these policies can be applied to their counterparts in ASEAN to help the latter understand their current performance and forecast their future positions in the world according to their planned goals.

References

Barrot, J.S. (2017). Research impact and productivity of Southeast Asian countries in language and linguistics. *Scientometrics,* 110 (1), 1-15.

Dodgson, M. (2002). Science, technology, and innovation: Issues and rationales. In M. Asher, D. Newman & T. Snyder (Eds.), *Public policy in Asia: Implications for business and government* (pp. 241-264). Westport: Quorum Books.

Erfanmanesh, M., Jahromi, R. B., Hosseini, E., & Gholamhosseinzadeh, Z. (2013). Scientific productivity, impact and collaboration of the top Asian countries in Scopus during 1996–2010. *COLLNET Journal of Scientometrics and Information Management*, 7(1), 97-110.

Kumar, S., & Jan, J.M. (2013). Mapping research collaborations in the business and management field in Malaysia, 1980-2010. *Scientometrics*, 97(3), 491-517.

Kumar, S., Rohani, V.A., & Ratnavelu, K. (2014). International research collaborations of ASEAN Nations in economics, 1979-2010. *Scientometrics*, 101(1), 847-867.

Lai, M., & Yap, S. (2004). Technology development in Malaysia and the newly industrializing economies: A comparative analysis. *Asia-Pacific Development Journal,* 11, 53-80.

Maula, A.W., Fuad, A., & Utarini, A. (2018). Ten-years trend of dengue research in Indonesia and South-east Asian countries: a bibliometric analysis. *Global Health Action*, 11(1), 1504398.

Navarrete, I.A., & Asio, V.B. (2014). Research productivity in soil science in the Philippines. *Scientometrics*, 100(1), 261-272.

Payumo, J.G., & Sutton, T.C. (2015). A bibliometric assessment of ASEAN collaboration in plant biotechnology. *Scientometrics*, 103(3), 1043-1059.

Prathap, G. (2018). A bibliometric tale of two cities: Hong Kong and Singapore. *Scientometrics*, 1-7. DOI: 10.1007/s11192-018-2927-z

Rana, S. (2012). Bibliometric analysis of output and visibility of science and technology in Singapore during 2000-2009. *Webology*, 9(1), 20-35.

Remøe, S. (2010). Supra-national governance of research and innovation policies and economic globalization: Tensions and dilemmas. *Journal of the Knowledge Economy,* 1(4), 268–288.

Rodriguez, V. and Soeparwata, A. (2012) ASEAN benchmarking in terms of science, technology, and innovation from 1999 to 2009**.** *Scientometrics*, 92(3), 549-573.

SciVal research metrics guidebook (2018). Retrieved from https://www.elsevier.com/__data/assets /pdf_file/0020/53327/ELSV-13013-Elsevier-Research-Metrics-Book-r5-Web.pdf (Accessed 21 November 2018).

Sigurdson, J., & Palonka, K. (2002). Technological Governance in ASEAN–Failings in technology transfer and domestic Research. Retrieved from https://pdfs.semanticscholar.org/ae2d/ fe92f23d991a4d50426b81323e78e74321f3.pdf (Accessed 21 November 2018).

Sombatsompop, N., Kositchaiyong, A., Markpin, T., & Inrit, S. (2006). Scientific evaluations of citation quality of international research articles in the SCI database: Thailand case study. *Scientometrics*, 66(3), 521-535.

Sombatsompop, N., Premkamolnetr, N., Markpin, T., Ittiritmeechai, S., Wongkaew, C., Yochai, W., ... & Beng, L.I. (2011). Viewpoints on synergising ASEAN academic visibilities through research

collaboration and the establishment of an ASEAN Citation Index Database. *Asia Pacific Viewpoint*, 52(2), 207-218.

Tahira, M., Abrizah, A., Alias, R.A. & Bakri, A. (2018). Quantitative evaluation of new research performance indices: A case of Malaysian engineering researchers. *Information Development*, 34(1), 64-76.

Tahira, M., Alias, R.A., Bakri, A. & Abrizah, A. (2016). Meso Level Institutional and Journal Related Indices for Malaysian Engineering Research. *Scientometrics* , 107(2), 521-535

Thomson Reuters. (2016). *ASEAN – The Emerging Research and Innovation Hub*. Bangkok: Thomson Reuters, Report.

Vinluan, L.R. (2012). Research productivity in education and psychology in the Philippines and comparison with ASEAN countries. *Scientometrics*, 91(1), 277-294.

Wagner C.S, Whetsell T, Baas J. and Jonkers, K (2018). Openness and impact of leading scientific countries. *Frontiers in. Research Metrics and Analytics,* 3 (10), doi: 10.3389/frma.2018.0001.

[1] Available at: https://www.scimagojr.com/countryrank.php?region=Asiatic%20Region (accessed 28 Dec 2018)

[2] ASEAN University Network, History and Background.Available at: http://www.aunsec.org/ourhistory.php

[3] www.aunsec.org

The convergent validity of several (field-normalized) bibliometric indicators: How well does *I3* perform for impact measurement?[1]

Lutz Bornmann[1], Alexander Tekles[2], Loet Leydesdorff[3]

[1] *bornmann@gv.mpg.de*
Division for Science and Innovation Studies, Administrative Headquarters of the Max Planck Society, Hofgartenstr. 8, 80539 Munich (Germany).

[2] *alexander.tekles.extern@gv.mpg.de*
Division for Science and Innovation Studies, Administrative Headquarters of the Max Planck Society, Hofgartenstr. 8, 80539 Munich (Germany);
Ludwig-Maximilians-University Munich, Institute of Sociology, Konradstr. 6, 80801 Munich (Germany).

[3] *loet@leydesdorff.net*
University of Amsterdam, P.O. Box 15793, 1001 NG Amsterdam (The Netherlands).

Abstract

Recently, the integrated impact indicator (*I3*) indicator was introduced where citations are weighted in accordance with the percentile rank class of each publication in a set of publications. *I3* can be used as a field-normalized indicator. Field-normalization is common practice in bibliometrics, especially when institutions and countries are compared. In this study, we test the ability of the indicator to discriminate between quality levels of papers as defined by Faculty members at F1000Prime. F1000Prime is a post-publication peer review system for assessing papers in the biomedical area. Thus, we test the convergent validity of *I3* (in its size-independent variant) using assessments by peers as baseline and compare its validity with several other (field-normalized) indicators: the mean-normalized citation score (MNCS), relative-citation ratio (RCR), citation score normalized by cited references (CSNCR), characteristic scores and scales (CSS), source-normalized citation score (SNCS), citation percentiles, and proportion of papers which belong to the *x*% most frequently cited papers ($PP_{top\ x\%}$). The results show that the $PP_{top\ 1\%}$ indicator discriminates best among different quality levels. *I3* performs similar as (slightly better than) most of the other field-normalized indicators. Thus, the results point out that the indicator could be a valuable alternative to other indicators in bibliometrics.

Introduction

In the application of citation analysis in research evaluation, one may need to compare the citation impact of publications from different fields. Different from using raw citation counts from the Web of Science (WoS, Clarivate Analytics) or Scopus (Elsevier) databases, professional bibliometricians have knowledge of differences in publication and citation cultures among fields of science (e.g., concerning the speed and frequency of citations) and use methods to assess the citation impact of focal papers against the impact of all other papers *in the same field and publication year* (McAllister, Narin, & Corrigan, 1983; Narin, 1981; Wang, Song, & Barabási, 2013). Field delineation, however, is not a *sine cure*.

Various indicators (approaches) have been introduced in bibliometrics since the early 1980s to construct field-normalized scores. According to Waltman (2016) "the idea of these indicators is to correct as much as possible for the effect of variables that one does not want to influence the outcomes of a citation analysis, such as the field, the year, and the document type of a publication" (p. 375). The necessity to normalize citation impact for cross-field comparisons is also one of the ten principles for research evaluation formulated in the Leiden Manifesto (Hicks, Wouters, Waltman, de Rijcke, & Rafols, 2015).

Leydesdorff and Bornmann (2011b) introduced the integrated impact indicator (*I3*) where citations are weighted in accordance with the percentile rank class of each publication in a set

[1] An extended version of the manuscript is in press at *Scientometrics* (Bornmann, Tekles, & Leydesdorff, in press).

of publications (e.g., published by a researcher or research group). Percentiles are *a priori* field-normalized: one can compare the top-1% for different reference sets. Although several publications appearing afterwards have dealt with the indicator (Leydesdorff & Bornmann, 2012; Rousseau, 2012; Wagner & Leydesdorff, 2012; Ye, Bornmann, & Leydesdorff, 2017), a comparison with other (field-normalized) indicators has not yet been done. In this study, therefore, we undertake this comparison by investigating the convergent validity of the indicator. In psychometrics, convergent validity tests whether measurements which are assumed to be related (here: assessments by peers and citation impact) are actually related: we are interested in the question of how *I3* discriminates between papers having received different quality scores by peers compared to various other indicators. We received a dataset from F1000Prime (see https://f1000.com/prime) including the bibliographic information of papers published in the biomedical area and their quality scores by peers. We use these scores as a benchmark for testing the indicators (Garfield, 1979).

Normalization of citation impact in bibliometrics

In this section, the various field-normalized indicators are explained which are used for the comparison with the *I3* indicator: mean-normalized citation score (MNCS), relative-citation ratio (RCR), citation score normalized by cited references (CSNCR), characteristic scores and scales (CSS), source normalized citation score (SNCS), citation percentiles, and proportion of papers which belong to the *x*% most frequently cited papers ($PP_{top\ x\%}$). More comprehensive overviews of methods for normalizing citations can be found in Mingers and Leydesdorff (2015) and Waltman (2016). *I3* is explained in the following after all other indicators have been explained, since the *I3* variant used in this study is based on other field-normalizing approaches. One can distinguish between field-normalization and statistical normalization: each indicator assumes some form of reference sets (field-normalization) and some form of comparison-strategy (statistical normalization). The indicators compared in this study vary with respect to both these aspects: different reference sets (e.g., papers published in the same subject category or co-cited papers) and different strategies to compare the focal papers to these reference sets (e.g., comparing values in relation to the mean or generating percentiles for comparison). Most of the variance among the selected indicators for this study result from statistical normalization. However, there is already (at least some) variance with respect to the choice of field categorization. Most of the indicators in this study have been calculated based on WoS subject categories. RCR and the citing side indicators are not relying on these categories, but on co-cited papers and papers published in the same journal.

The use of WoS subject categories for field-normalization has been criticized, because WCs are attributed to journals (and not to individual papers) and journals are not homogeneous in terms of the disciplines of papers published in them (Leydesdorff & Bornmann, 2016). Although other field-categorisation schemes have been proposed for the normalization of citation impact such as algorithmically constructed classification systems (Ruiz-Castillo & Waltman, 2015) or expert-based field categorisations (Bornmann, Marx, & Barth, 2013) "the WoS journal subject categories are the most commonly used field classification system for normalisation purposes" (Wouters et al., 2015, p. 18).

All indicators considered here with the exception of the RCR are available at the paper level in an in-house database of the Max Planck Society which is based on the Web of Science (WoS, Clarivate Analytics). We retrieved the additional RCR scores in a two-step process. First, the DOIs are used to automatically query the papers' Pubmed IDs using the web form available under https://icite.od.nih.gov/analysis. Each of these requests returns an HTML document containing the Pubmed ID of the corresponding papers. Second, the Pubmed IDs were extracted from the HTML documents and used for requesting the papers' RCR scores via the iCite API at https://icite.od.nih.gov/api.

Mean-normalized citation score (MNCS)

Based on early proposals by Schubert and Braun (1986), the MNCS (Waltman, van Eck, van Leeuwen, Visser, & van Raan, 2011a) is calculated by dividing the citations of a paper in question by the average citation rate of the papers that were published in the same subject category (and publication year).

Relative-citation ratio (RCR)

Hutchins, Yuan, Anderson, and Santangelo (2016) proposed the Relative Citation Rate (RCR) as a new field-normalized impact indicator. The indicator is similarly designed as MNCS: it is a quotient of the focal paper's citation counts and the expected number of citations in the reference set. The difference of the RCR from the MNCS is that the expected value (respectively the reference sets) is based on co-citations: the papers co-cited with the focal paper are considered to represent a more precise reference set at the paper level than WCs which are attributed at the journal level. In bibliometrics, co-citations are frequently used similarity measures which are based on citation relations. An overview of research on the RCR can be found in Lindner, Torralba, and Khan (2018).

Citation score normalized by cited references (CSNCR)

Bornmann and Haunschild (2016) introduced the field-normalized indicator "citation score normalized by cited references" (CSNCR) which is closely related to the MNCS. The indicator is rooted in early suggestions by Garfield (1979) that "the most accurate measure of citation potential is the average number of references per paper published in a given field". The CSNCR is defined as follows: the citations of a focal paper are divided by the mean number of cited references in a subject category. The theoretical analysis of the CSNCR by Bornmann and Haunschild (2016) demonstrated that the indicator has the properties of consistency and homogeneous normalization. The authors' empirical comparison of the CSNCR with other field-normalized indicators revealed that it is as suitable as other field-normalized indicators to normalize citations.

Characteristic scores and scales (CSS)

The characteristic scores and scales (CSS) method by Glänzel and Schubert (1988) for normalizing citation data is one of the earliest proposed field-normalization approaches. The CSS method classifies the publications in reference sets (subject categories) as follows: "characteristic scores are obtained from iteratively truncating a distribution according to conditional mean values from the low end up to the high end. In particular, the scores b_k ($k > 0$) are obtained from iteratively truncating samples at their mean value and recalculating the mean of the truncated sample until the procedure is stopped or no new scores are obtained" (Glänzel, 2013, p. 111). In many studies based on this method, four impact classes are used to group the papers in reference sets (see Glänzel, Thijs, & Debackere, 2014):

1. poorly cited (papers with less citations than b_1),
2. fairly cited (papers with citations above b_1 but less citations than b_2),
3. remarkably cited (papers with citations above b_2 but less citations than b_3), and
4. outstandingly cited (papers with citations of at least b_4).

In the MPG in-house database, all papers in each reference set published since 1980 are classified following the CSS method.

Citing-side normalization of citation impact

Citations are attributed to papers on the cited side by the indicators mentioned above. Zitt and Small (2008) first introduced the idea of normalizing citation impact on the citing-side (van Leeuwen, Visser, Moed, Nederhof, & van Raan, 2003, already added the impact of citing

journals to the link between paper and received citation, thereby weighting the otherwise value of 1, by the impact score of the citing journal). The authors proposed a modification of the journal impact factor (JIF) by fractional citation weighting. Citing-side normalization is also named source normalization, fractional citation weighting, fractional counting of citations, or a priori normalization (Waltman & van Eck, 2013a). The method cannot only be used for journals as initially proposed by Zitt and Small (2008) but also for any other publication sets (Moed, 2010). Citing-side normalization considers the environment of a given citation (Leydesdorff & Bornmann, 2011a; Leydesdorff, Radicchi, Bornmann, Castellano, & de Nooy, 2013): the citation is weighted depending on its environment. A citation from a subject category with papers containing long reference lists (e.g., bio-medicine) receives a lower weighting than a citation from a subject category with on average only few citations.
For citing-side normalization, the number of references of the citing paper is usually used to weight a specific citation (Waltman & van Eck, 2013b). The assumption is that this number of references reflects the typical number in the field (subject categories) of the citing paper. However, this assumption cannot always be made. For this reason, an average number of references is calculated (and used as weighting factor) which includes other papers appearing in a journal alongside the citing paper. In this study, we consider three variants of citing-side normalization, which are explained by Waltman and van Eck (2013b) in more detail: SNCS1, SNCS2, and SNCS3 (source normalized citation score).

Percentile-based indicators

Citation impact percentiles

The distribution of citation data is usually very skewed with only a few papers being highly-cited (Seglen, 1992). Since the arithmetic mean is not appropriate as a measure of the central tendency in a skewed distribution, citation impact percentiles have been introduced as an alternative to approaches based on the averages of citations. The citation impact percentile of a specific paper indicates the share of other papers in the reference set which have received fewer citations. For example, a citation impact percentile of 80 indicates that 80% of the papers in the reference set have received fewer citations.
Citation impact percentiles from different reference sets are directly comparable with one another; no further field-normalization is needed. Suppose the citation impact of two papers have been normalized based on different reference sets and both papers have a percentile of 70. The identical percentile indicates that both papers have – compared with the other papers in the corresponding reference sets – achieved the same citation impact. Even though both papers may have different times cited values in the WoS database, the relative citation impacts are the same.
Citation impact percentiles can be calculated with various procedures (see the overview in Bornmann, Leydesdorff, & Mutz, 2013). In the current study, two approaches were used which are frequently applied in evaluative bibliometrics. For both approaches, all papers in the reference sets are ranked in decreasing or increasing order by their citation counts (i), and the number of publications in the reference set (n) is determined in the first step. For the product InCites (a customised, web-based research evaluation tool based on bibliometric data from WoS), Clarivate Analytics calculates the percentiles by using (basically) the formula ($[i/n] * 100$). This inversed ranking will be named as "InCites percentiles" in the following. However, the use of this formula may lead to a mean percentile of a reference set unequal to 50 (the median). The formula ($[(i - 0.5)/n] * 100$) (Hazen, 1914) does not suffer this disadvantage. We will use the abbreviation "Hazen percentile" for these percentiles in the following. Furthermore, the papers are sorted in increasing impact order for InCites percentiles, but in decreasing order for Hazen percentiles; we invert the InCites percentiles in this study by subtracting the values from 100.

Proportion of papers belonging to the top-*x*%

Citation percentiles can be directly used for impact measurements. However, it is also very common in bibliometrics to focus on certain percentile classes (Bornmann, 2014). In this study, we include three indicators focusing on three classes: $PP_{top\text{-}50\%}$, $PP_{top\text{-}10\%}$, and $PP_{top\text{-}1\%}$. The indicators reveal the proportion of papers published by a unit which belong to the *x*% most frequently cited papers. The results of Tahamtan and Bornmann (2018) show that the $PP_{top\text{-}x\%}$ indicators – especially the $PP_{top\text{-}10\%}$ indicator – are one of the earliest used field-normalized indicators in scientometrics which were introduced by Narin (1981). Tijssen, Visser, and van Leeuwen (2002) discuss pros and cons of using $PP_{top\text{-}x\%}$ indicators. In this study, we used $PP_{top\text{-}x\%}$ indicators which have been calculated based on two fractional counting approaches.

Papers may be equal in the rankings, if the papers are sorted by citations and more than one paper have the same citation counts. These ties in citations lead to the problem of exactly assigning the papers to the top-*x*% class or the corresponding bottom-*x*% class. To solve this problem we use an approach introduced by Waltman and Schreiber (2013). They propose to fractionally assign the papers at the top-*x*% threshold to the top- and bottom-*x*% – in dependence of the number of papers with the same number of citations at the threshold. The second fractional counting approach used for the indicators concerns the multiple assignment of journals to subject categories. We use the fractional counting approach by Waltman, van Eck, van Leeuwen, Visser, and van Raan (2011b) to calculate the $PP_{top\text{-}x\%}$ indicators across multiple subject categories.

I3 indicator

One of the newest (field-normalized) indicators is the *I3* indicator which is also a percentile-based indicator. It was defined as a non-parametric alternative to field-normalized indicators based on mean citations. Bornmann (2010) and Bornmann and Mutz (2011) proposed to use the weighted number of papers of units (e.g., journals or institutions) belonging to certain percentile impact classes for performance measurements. The further elaboration into *I3*, the integrated impact indicator, combines these proposals in a unified scheme (Leydesdorff & Bornmann, 2011b; Leydesdorff & Bornmann, 2012; Rousseau, 2012; Wagner & Leydesdorff, 2012).

I3 weights citations in accordance with the percentile rank class of each paper in a set of papers. Since the paper numbers in the higher top-*x*% classes are subsets of the numbers in the lower classes, the numbers in the percentile classes have been corrected correspondingly to avoid double counting of papers. In the most recent development, Leydesdorff, Bornmann, and Adams (in press) propose to use four percentile classes (top-1%, top-10%, top-50%, and bottom-50%) as weighting scheme for *I3*. They argued that a paper in the top-1% class can be valued ten times more than a paper in the top-10% class. It follows that a top-1% paper weights 100 times a paper at the bottom. It is an advantage of this scheme that it appreciates the highly-skewed nature of citation data by using a logarithmic scale. It follows that papers in the top-50% are weighted with two and bottom-50% with one. The resulting indicator correlates above .9 with the numbers of both publications and citations in empirical cases.

Methods

Peer ratings provided by F1000Prime

F1000Prime is a post-publication peer review system of papers published in medical and biological journals. The service started with F1000 Biology in 2002; F1000 Medicine followed in 2006. Both services were merged in 2009 to the current F1000Prime database. Papers which are included in the F1000Prime database are selected by a peer-nominated global "Faculty". These are leading scientists and clinicians who assess the papers and explain their importance.

F1000Prime covers a restricted set of papers published in medical and biological journals; most of the papers from these journals are not selected (Kreiman & Maunsell, 2011; Wouters & Costas, 2012).
The Faculty includes more than 5,000 experts worldwide. Faculty members can choose and assess any paper of interest. Although many papers published in popular and reputable journals (e.g., *Nature* and *Science*) are evaluated by the members, most of the papers have been published in specialised or less well-known journals (Wouters & Costas, 2012). "Less than 18 months since Faculty of 1000 was launched, the reaction from scientists has been such that two-thirds of top institutions worldwide already subscribe, and it was the recipient of the Association of Learned and Professional Society Publishers (ALPSP) award for Publishing Innovation in 2002 (http://www.alpsp.org/about.htm)" (Wets, Weedon, & Velterop, 2003, p. 249).
The selected papers for F1000Prime are rated by the Faculty members as "good", "very good", or "exceptional" which are set to the scores of 1, 2, or 3, respectively. Since many papers are assessed not only by one Faculty member but by several, we calculated the sum of the scores for this study. This accords to the F1000Prime practice to use the individual scores for calculating the total score for each paper (which are used then to rank the papers in the disciplines). The assessments in the F1000 database can be used either by scientists for receiving pointers to relevant papers in their areas, but also as a database for research evaluation purposes. According to Wouters and Costas (2012) "the data and indicators provided by F1000 are without doubt rich and valuable, and the tool has a strong potential for research evaluation, being in fact a good complement to alternative metrics for research assessments at different levels (papers, individuals, journals, etc.)" (p. 14).

Used datasets

In 2018, F1000 provided one of the authors with data on recommendations made by the Faculty members and the bibliographic information for the corresponding papers in their system (n=51,461 papers). We matched the papers with the papers in our WoS in-house database (of the Max Planck Society) using the DOI. We restricted the set to papers with the document types "article" and "review". In the statistical analyses, we included not only the field-normalized indicators explained above (with a citation window between publication year and the end of 2017), but also citation counts (1) for a three-year citation window and (2) for the period between publication year and the end of 2017. We included only matched F1000Prime papers into the study until 2015 to ensure a minimum citation window of three years (Glänzel & Schöpflin, 1995). Since the indicators are concerned by different numbers of missing values, only papers have been considered with no missing value across all indicators. These restrictions lead to a total number of 28,063 papers for the statistical analysis published between 2000 and 2015 (most of the reduction is due to the necessity of using a minimum citation window).

Results

Since *I3* can be used as field-normalized indicator, we are interested in how it discriminates between papers rated differently by Faculty members compared to other (field-normalized) indicators. In other words, we are interested in its convergent validity: does the indicator discriminate worse, equal to, or better than the other indicators between the different quality levels and is thus more convergently valid to the assessment by peers than the other indicators? *I3* differs from the other indicators by being calculated on the aggregated, and not on the single paper level. Thus, we need groups of papers for the comparison of *I3* with other indicators.
The CSS method which we explained above cannot only be used to field-normalize single papers, but to group any paper set with metrics (see, e.g., Bornmann & Glänzel, 2018). Using the CSS method to group the papers in four classes – based on the sum of the F1000Prime scores per paper – we found 1396 papers (4.97%) in the class with the best scores (F1000 class

4, sum scores between 5 and 35), 3737 papers (13.32%) in the second best class (F1000 class 3, sum scores between 3 and 4), 10,334 papers (36.82%) in the next class (F1000 class 2, sum scores equal to 2), and 12,596 papers (44.88%) in the lowest class (F1000 class 1, sum scores equal to 1).

For the four groups, we calculated the arithmetic average of each indicator per group. The median would have been an alternative, but this statistic fails to properly differentiate between the groups because of ties in certain indicator values. For example, the $PP_{top\ x\%}$ indicators mostly consists of the values 0 and 1 which lead to corresponding indifferent median values for the classes. We decided not to use the sum, since the results are dependent on the sample size: the more papers in a group are, the better results can be expected.

Although *I3* was designed to reflect the output in addition to the impact dimension (as a sum score), the output dimension is not relevant for this validity study. The performance of the four F1000 classes is not dependent on the output dimension; only the impact of the single papers matters. In the usual evaluation of research groups or institutions, however, we are faced with a different situation in which both dimensions – publications and citations – are of equal interest for assessing performance.

In case of the *I3* indicator, we divided *I3* by the number of papers in a group and obtain *I3*/*N*. For the four F1000Prime quality groups, we received the following *I3*/*N* values: F1000 class 1 = 11.68, F1000 class 2 = 14.63, F1000 class 3 = 22.66, and F1000 class 4 = 39.03. The mean values point out that *I3* measures quality as expected: it discriminates validly between the four performance groups. However, does *I3/N* discriminate better between the groups than the other indicators (and is thus more convergently valid)? As the results in Table 1 show, all other indicators which we considered in this study are similarly able to discriminate between the four F1000 classes.

Table 1. Mean indicator scores for four F1000 classes (class 4 reflects the highest quality level).

Indicator (mean value)	F1000 class			
	1 (*n*=12,596)	2 (*n*=10,334)	3 (*n*=3737)	4 (*n*=1396)
$PP_{top\ 50\%}$	0.87	0.91	0.96	1.00
$PP_{top\ 10\%}$	0.41	0.50	0.68	0.89
$PP_{top\ 1\%}$	0.07	0.10	0.17	0.33
Number of citations (until 2017)	54.03	58.77	93.92	157.69
Number of citations (three year citation window)	31.08	38.62	60.42	115.69
MNCS	3.18	3.68	5.60	9.92
CSNCR	4.19	4.79	7.59	13.68
CSS	0.59	0.64	0.89	1.24
SNCS1	3.57	4.00	6.07	10.71
SNCS2	3.14	3.53	5.28	9.18
SNCS3	3.30	3.71	5.54	9.57
Hazen percentiles	78.43	82.92	89.25	95.93
Incites percentiles	78.27	82.76	89.27	95.84
RCR	3.72	4.15	6.42	11.54
I3	11.68	14.63	22.66	39.03

To compare the ability of the indicators to discriminate between the four F1000 classes, we calculated the so called "Average Annual Growth Rate (AAGR)" (instead of annual differences we have quality group differences in our study). The AAGR is the average increase in citation impact over the quality groups. It is computed by taking the arithmetic average of a series of growth rates. In the first step of calculating AAGR for each indicator, we determined the percentage growth for each group (except for F1000 class 1) which is the percentage growth (F1000 class x / F1000 class $x - 1$) – 1. In the second step, the AAGR is calculated as the sum of each indicator's growth rate divided by the number of F1000 classes – 1. We also calculated the "Sum Annual Growth Rate (SAGR) for comparison with the AAGR which is a measure of the total increase in citation impact over the quality groups.

Table 2. AAGR and SAGR for the various indicators. The indicators are ordered by SAGR (and AAGR) in decreasing order. The column "Difference to previous SAGR" shows how much the SAGR of an indicator differs from its previous SAGR with the rank $x - 1$.

Indicator (mean value)	AAGR	SAGR	Rank	Difference to previous SAGR
$PP_{top\ 1\%}$	67.24	201.72	1	
Number of citations (three year citation window)	57.39	172.17	2	-29.55
CSNCR	50.99	152.98	3	-19.19
I3	50.79	152.38	4	-0.60
RCR	48.70	146.10	5	-6.28
MNCS	48.35	145.04	6	-1.06
SNCS1	46.67	140.01	7	-5.03
Number of citations (until 2017)	45.50	136.49	8	-3.52
SNCS2	45.26	135.78	9	-0.71
SNCS3	44.80	134.41	10	-1.37
$PP_{top\ 10\%}$	29.52	88.57	11	-45.84
CSS	29.16	87.48	12	-1.09
Incites percentiles	6.99	20.97	13	-66.51
Hazen percentiles	6.95	20.84	14	-0.12
$PP_{top\ 50\%}$	4.49	13.48	15	-7.37

The results on the basis of AAGR and SAGR for the various indicators are shown in Table 2. The indicators are sorted by SAGR (and AAGR) in decreasing order. The column "Difference to previous SAGR" reveals how much the SAGR of an indicator differs from the SAGR of the indicator with the rank $x - 1$. Thus, the column indicates how much larger the scores in the better class are. The results in Table 2 point out that $PP_{top\ 1\%}$ discriminates best between the different quality classes. The indicator is followed by the number of citations (measured across a three year citation window). CSNCR is on the third position whereby *I3* has very similar AAGR and SAGR values as CSNCR.

As the "Difference to previous SAGR" column reveals, $PP_{top\ 1\%}$ discriminates much better than the second best positioned indicator number of citations (measured across a three year citation window) which performs itself much better than the CSNCR indicator. The indicators with the rank positions 3 to 10 are able to discriminate similarly between the four quality levels. The next larger performance difference are visible between $PP_{top\ 10\%}$ and SNCS3 (-45.84%) as well as between Incites percentiles and CSS (-66.51%).

Discussion

The discussion about the normalization of citation impact has a long tradition in bibliometrics. Since publication and citation practices are very different among the various fields of science, citation numbers from different fields cannot be directly compared. The use of field-normalized indicators in research evaluation is one guiding principle in the Leiden Manifesto (Hicks et al., 2015). The same Manifesto advocates the use of percentiles for field normalization.

In many evaluation contexts one uses field-normalized indicators (based on statistical normalization by the mean) for measuring citation impact instead of using the raw times-cited information from the WoS or Scopus databases. For example, field-normalized indicators are used in the popular Times Higher Education Rankings (see https://www.timeshighereducation.com/world-university-rankings).

Research on these indicators focused especially on the use of the arithmetic average of highly-skewed citation distributions. This poses a problem, for instance, for the use of MNCS and the way in which "research fields" are operationalized. Various categorization schemes can be used to define fields (e.g., schemes based on citation relations or subject categorizations from field-specific literature databases) and fields can be defined at different levels of aggregation (Wilsdon et al., 2015).

Some research has been undertaken hitherto to identify field-normalized indicators using methods which normalize citation impact better than other indicators. According to the empirical results of Waltman and van Eck (2013b), citing-side normalization has been shown more successful than cited-side normalization in field-normalizing citation impact. Purkayasthaa, Palmaroa, Falk-Krzesinskib, and Baas (2018) reported the following results: "from the high correlations within our analyses of the two metrics across a range of research areas, we conclude that RCR_{Scopus} and FWCI [field-weighted citation impact] can be used interchangeably to evaluate citation impact of an article or of larger entities such as universities". Bornmann and Leydesdorff (2013) and Bornmann and Marx (2015) used assessments from F1000Prime to compare the validity of different (field-normalized) citation impact indicators.

We included a range of (field-normalized) indicators in the current study to compare the newly proposed *I3* indicator with other indicators with respect to their convergent validity (using assessments by peers as a baseline; sometimes called the "gold standard" of peer review). We wanted to know whether *I3* is better able than other indicators in discriminating between different quality levels as defined by Faculty members working for F1000Prime. The indicators differ in terms of field-categorization (e.g., papers in the same WoS subject category or co-cited papers) and comparison-strategy (e.g., comparison of percentiles or focal papers with mean values). The investigation of the different indicators show smaller differences between different types of reference sets (field-categorization), but larger differences with respect to the comparison strategy (statistical normalization).

The results show that the $PP_{top\ 1\%}$ indicator discriminates best compared to the other indicators given the assumed baseline of F1000Prime. However, this result reflects the orientation of F1000Prime towards excellence in biomedicine which the $PP_{top\ 1\%}$ indicator targets more precisely than any of the other indicators. The second best indicator is the raw number of citations in the first three years after publication. Although this indicator is not field-normalized nor statistically normalized, it performs comparably well – perhaps because it focusses specifically on the period when most of the papers are selected by the Faculty members for inclusion in the F1000Prime database. Perhaps, the Faculty members also consider the number of citations in their selection decisions and assessments of the papers. Furthermore, the F1000Prime dataset is a relatively homogenous dataset with respect to field differences, and for this reason field-normalization may not play an important role.

At the third and fourth positions in the validity ranking of the indicators are CSNCR and *I3* with a very similar value. Both indicators also differ scarcely from (perform slightly better than) RCR, MNCS, and the three SNSCI indicators (as well as the citation counts measured over the variable citation window until 2017). Thus, the newly developed *I3* indicator holds up well against many other (field-normalized) indicators by discriminating equal to (or even slightly better than) the other indicators between the four F1000 quality classes.

With regard to percentiles (InCites and Hazen percentiles), our results are in disagreement to the previous results of Bornmann and Leydesdorff (2013). They reported very positive results for citation percentiles when this indicator is compared with other (field-normalized) indicators: "Percentile in Subject Area achieves the highest correlation with F1000 ratings" (p. 286). Using other data, the results in this study show, however, that percentiles (InCites and Hazen percentiles) perform comparably worse. The reasons for the differences between both studies should further be investigated in future studies.

A reason for the comparably poor performance of some of the percentile-based indicators might be that, even in the already selective F1000 dataset, highly skewed distributions of quality scores and indicator values can be observed. Most of the papers fall into F1000 classes 1 or 2 which are very similar when compared to indicators which include low quality scores. This is also reflected in the indicator values across classes: for most of the indicators, classes 1 and 2 are rather similar, whereas class 4 substantially differs from the other classes. As a result, the assessment of the indicators' validity mainly rests on the ability to discriminate the top papers from the rest of the (already selective set of) papers. This may also favor percentile-based indicators focusing on the upper end of the citation distribution. We expect that other percentile-based indicators would be better able to differentiate between papers (groups of papers) reflecting the broad range of different quality levels.

Acknowledgments

The bibliometric data used in this paper are from an in-house database developed and maintained in cooperation with the Max Planck Digital Library (MPDL, Munich) and derived from the Science Citation Index Expanded (SCI-E), Social Sciences Citation Index (SSCI), Arts and Humanities Citation Index (AHCI) prepared by Clarivate Analytics, formerly the IP & Science business of Thomson Reuters (Philadelphia, Pennsylvania, USA). We would like to thank Tom Des Forges and Ros Dignon from F1000 for providing us with the F1000Prime dataset.

References

Bornmann, L. (2010). Towards an ideal method of measuring research performance: some comments to the Opthof and Leydesdorff (2010) paper. *Journal of Inormetrics, 4*(3), 441-443.

Bornmann, L. (2014). How are excellent (highly cited) papers defined in bibliometrics? A quantitative analysis of the literature. *Research Evaluation, 23*(2), 166–173.

Bornmann, L., & Glänzel, W. (2018). Which differences can be expected when two universities in the Leiden Ranking are compared? Some benchmarks for institutional research evaluations. *Scientometrics, 115*(2), 1101-1105. doi: 10.1007/s11192-018-2700-3.

Bornmann, L., & Haunschild, R. (2016). Citation score normalized by cited references (CSNCR): The introduction of a new citation impact indicator. *Journal of Informetrics, 10*(3), 875-887.

Bornmann, L., & Leydesdorff, L. (2013). The validation of (advanced) bibliometric indicators through peer assessments: A comparative study using data from InCites and F1000. *Journal of Informetrics, 7*(2), 286-291. doi: 10.1016/j.joi.2012.12.003.

Bornmann, L., Leydesdorff, L., & Mutz, R. (2013). The use of percentiles and percentile rank classes in the analysis of bibliometric data: opportunities and limits. *Journal of Informetrics, 7*(1), 158-165. doi: 10.1016/j.joi.2012.10.001.

Bornmann, L., & Marx, W. (2015). Methods for the generation of normalized citation impact scores in bibliometrics: Which method best reflects the judgements of experts? *Journal of Informetrics, 9*(2), 408-418.

Bornmann, L., Marx, W., & Barth, A. (2013). The normalization of citation counts based on classification systems. *Publications, 1*(2), 78-86.

Bornmann, L., & Mutz, R. (2011). Further steps towards an ideal method of measuring citation performance: The avoidance of citation (ratio) averages in field-normalization. *Journal of Informetrics, 5*(1), 228-230. doi: 10.1016/j.joi.2010.10.009.

Bornmann, L., Tekles, A., & Leydesdorff, L. (in press). How well does I3 perform for impact measurement compared to other bibliometric indicators? The convergent validity of several (field-normalized) indicators. *Scientometrics*.

Garfield, E. (1979). *Citation indexing - its theory and application in science, technology, and humanities*. New York, NY, USA: John Wiley & Sons, Ltd.

Glänzel, W. (2013). The application of citation-based performance classes to the disciplinary and multidisciplinary assessment in national comparison. In J. Gorraiz, E. Schiebel, C. Gumpenberger, M. Hörlesberger & H. F. Moed (Eds.), *Proceedings of ISSI 2013 - 14th International Society of Scientometrics and Informetrics Conference* (Vol. 1, pp. 109-122). Vienna.

Glänzel, W., & Schöpflin, U. (1995). A Bibliometric Study on Aging and Reception Processes of Scientific Literature. *Journal of Information Science, 21*(1), 37-53. doi: 10.1177/016555159502100104.

Glänzel, W., & Schubert, A. (1988). Characteristic scores and scales in assessing citation impact. *Journal of Information Science, 14*(2), 123-127.

Glänzel, W., Thijs, B., & Debackere, K. (2014). The application of citation-based performance classes to the disciplinary and multidisciplinary assessment in national comparison and institutional research assessment. *Scientometrics, 101*(2), 939-952. doi: 10.1007/s11192-014-1247-1.

Hazen, A. (1914). Storage to be provided in impounding reservoirs for municipal water supply. *Transactions of American Society of Civil Engineers, 77*, 1539-1640.

Hicks, D., Wouters, P., Waltman, L., de Rijcke, S., & Rafols, I. (2015). Bibliometrics: The Leiden Manifesto for research metrics. *Nature, 520*(7548), 429-431.

Hutchins, B. I., Yuan, X., Anderson, J. M., & Santangelo, G. M. (2016). Relative Citation Ratio (RCR): A new metric that uses citation rates to measure influence at the article level. *Plos Biology, 14*(9), e1002541.

Kreiman, G., & Maunsell, J. H. R. (2011). Nine criteria for a measure of scientific output. *Frontiers in Computational Neuroscience, 5*. doi: 10.3389/fncom.2011.00048.

Leydesdorff, L., & Bornmann, L. (2011a). How fractional counting of citations affects the Impact Factor: normalization in terms of differences in citation potentials among fields of science. *Journal of the American Society for Information Science and Technology, 62*(2), 217-229. doi: 10.1002/asi.21450.

Leydesdorff, L., & Bornmann, L. (2011b). Integrated Impact Indicators (I3) compared with Impact Factors (IFs): an alternative research design with policy implications. *Journal of the American Society of Information Science and Technology, 62*(11), 2133-2146.

Leydesdorff, L., & Bornmann, L. (2012). Percentile ranks and the integrated impact indicator (I3). *Journal of the American Society For Information Science and Technology, 63*(9), 1901-1902. doi: 10.1002/asi.22641.

Leydesdorff, L., & Bornmann, L. (2016). The operationalization of "fields" as WoS subject categories (WCs) in evaluative bibliometrics: The cases of "library and information science" and "science & technology studies". *Journal of the Association for Information Science and Technology, 67*(3), 707-714. doi: 10.1002/asi.23408.

Leydesdorff, L., Bornmann, L., & Adams, J. (in press). The Integrated Impact Indicator (I3) Revisited: A Non-Parametric Alternative to the Journal Impact Factor. *Scientometrics, 2018*.

Leydesdorff, L., Radicchi, F., Bornmann, L., Castellano, C., & de Nooy, W. (2013). Field-normalized impact factors: a comparison of rescaling versus fractionally counted IFs. *Journal of the American Society for Information Science and Technology, 64*(11), 2299-2309.

Lindner, M. D., Torralba, K. D., & Khan, N. A. (2018). Scientific productivity: An exploratory study of metrics and incentives. *PLOS ONE, 13*(4), e0195321. doi: 10.1371/journal.pone.0195321.

McAllister, P. R., Narin, F., & Corrigan, J. G. (1983). Programmatic evaluation and comparison based on standardized citation scores. *IEEE Transactions on Engineering Management, 30*(4), 205-211.

Mingers, J., & Leydesdorff, L. (2015). A Review of Theory and Practice in Scientometrics. *European Journal of Operational Research, 246*(1), 1-19.

Moed, H. F. (2010). Measuring contextual citation impact of scientific journals. *Journal of Informetrics, 4*(3), 265-277. doi: 10.1016/j.joi.2010.01.002.

Narin, F. (1981). *Concordance between subjective and bibliometric indicators of the nature and quality of performed biomedical research.* Washington DC, USA: US Department of Health and Human Services.

Purkayasthaa, A., Palmaroa, E., Falk-Krzesinskib, H. J., & Baas, J. (2018). Comparison of two article-level, feld-independent citation metrics: Field-Weighted Citation Impact (FWCI) and Relative Citation Ratio (RCR). Retrieved April 1, 2019, from http://dx.doi.org/10.2139/ssrn.3237564

Rousseau, R. (2012). Basic properties of both percentile rank scores and the I3 indicator. *Journal of the American Society for Information Science and Technology, 63*(2), 416-420. doi: 10.1002/asi.21684.

Ruiz-Castillo, J., & Waltman, L. (2015). Field-normalized citation impact indicators using algorithmically constructed classification systems of science. *Journal of Informetrics, 9*(1), 102-117. doi: 10.1016/j.joi.2014.11.010.

Schubert, A., & Braun, T. (1986). Relative indicators and relational charts for comparative assessment of publication output and citation impact. *Scientometrics, 9*(5-6), 281-291.

Seglen, P. O. (1992). The skewness of science. *Journal of the American Society for Information Science, 43*(9), 628-638.

Tahamtan, I., & Bornmann, L. (2018). Creativity in Science and the Link to Cited References: Is the Creative Potential of Papers Reflected in their Cited References? *Journal of Informetrics, 12*(3), 906-930.

Tijssen, R., Visser, M., & van Leeuwen, T. (2002). Benchmarking international scientific excellence: are highly cited research papers an appropriate frame of reference? *Scientometrics, 54*(3), 381-397.

van Leeuwen, T. N., Visser, M. S., Moed, H. F., Nederhof, T. J., & van Raan, A. F. J. (2003). Holy Grail of science policy: exploring and combining bibliometric tools in search of scientific excellence. *Scientometrics, 57*(2), 257-280.

Wagner, C. S., & Leydesdorff, L. (2012). An Integrated Impact Indicator: A new definition of 'Impact' with policy relevance. *Research Evaluation, 21*(3), 183–188. doi: 10.1093/reseval/rvs012.

Waltman, L. (2016). A review of the literature on citation impact indicators. *Journal of Informetrics, 10*(2), 365-391.

Waltman, L., & Schreiber, M. (2013). On the calculation of percentile-based bibliometric indicators. *Journal of the American Society for Information Science and Technology, 64*(2), 372-379.

Waltman, L., van Eck, N., van Leeuwen, T., Visser, M., & van Raan, A. (2011a). Towards a new crown indicator: an empirical analysis. *Scientometrics, 87*(3), 467-481. doi: 10.1007/s11192-011-0354-5.

Waltman, L., & van Eck, N. J. (2013a). Source normalized indicators of citation impact: an overview of different approaches and an empirical comparison. *Scientometrics, 96*(3), 699-716. doi: 10.1007/s11192-012-0913-4.

Waltman, L., & van Eck, N. J. (2013b). A systematic empirical comparison of different approaches for normalizing citation impact indicators. *Journal of Informetrics, 7*(4), 833-849.

Waltman, L., van Eck, N. J., van Leeuwen, T. N., Visser, M. S., & van Raan, A. F. J. (2011b). Towards a new crown indicator: some theoretical considerations. *Journal of Informetrics, 5*(1), 37-47. doi: 10.1016/j.joi.2010.08.001.

Wang, D., Song, C., & Barabási, A.-L. (2013). Quantifying Long-Term Scientific Impact. *Science, 342*(6154), 127-132. doi: 10.1126/science.1237825.

Wets, K., Weedon, D., & Velterop, J. (2003). Post-publication filtering and evaluation: Faculty of 1000. *Learned Publishing, 16*(4), 249-258.

Wilsdon, J., Allen, L., Belfiore, E., Campbell, P., Curry, S., Hill, S., . . . Johnson, B. (2015). *The Metric Tide: Report of the Independent Review of the Role of Metrics in Research Assessment and Management*. Bristol, UK: Higher Education Funding Council for England (HEFCE).

Wouters, P., & Costas, R. (2012). *Users, narcissism and control – tracking the impact of scholarly publications in the 21st century*. Utrecht, The Netherlands: SURFfoundation.

Wouters, P., Thelwall, M., Kousha, K., Waltman, L., de Rijcke, S., Rushforth, A., & Franssen, T. (2015). *The Metric Tide: Literature Review (Supplementary Report I to the Independent Review of the Role of Metrics in Research Assessment and Management)*. London, UK: Higher Education Funding Council for England (HEFCE).

Ye, F. Y., Bornmann, L., & Leydesdorff, L. (2017). h-based I3-type Multivariate Vectors: Multidimensional Indicators of Publication and Citation Scores. *COLLNET Journal of Scientometrics and Information Management, 11*(1), 153-171.

Zitt, M., & Small, H. (2008). Modifying the journal impact factor by fractional citation weighting: The audience factor. *Journal of the American Society for Information Science and Technology, 59*(11), 1856-1860. doi: 10.1002/asi.20880.

Technology Opportunity Analysis of Internet of Things From the Perspective of "Technology-Market"

Liu Jianhua[1]，Pan Song[1]，Li Yafei[2],Jiang Zhaohua[2]
（1. School of Management Engineering，Zhengzhou University，Zhengzhou 450001, China；
2.Institute of Science and Technology Management，Dalian University of Technology，Dalian Liaoning 116023, China）

Abstract： The Internet of Things is an important part of the new generation of information technology. Technology innovation and market exploration are critical ways for its industrial innovation. Based on Naoki Shibata's technology opportunity identification theory, the front-end and mid-end technology opportunities of the Internet of Things are identified by detecting and comparing the burst words of patents and scientific papers. Through the changes of the number of patents in the same family, the market opportunities of the Internet of Things mid-end technology are verified. Through the analysis of patent overseas layout, this paper shows the technological development of countries in the field of Internet of Things, and further identifies potential markets with great appeal on a global scale. Finally, this paper summarizes the meanings of the research on the technology innovation and market development of Internet of Things in China.

Key words： Internet of Things; technology opportunities; burst detection; patent layout

1 Introduction

The term "Internet of Things" was first proposed by Professor Ashton of MIT when he studied RFID technology in 1999[1]. The Internet of Things (IOT) is a network in which sensors exchange information and communicate with each other in accordance with the agreed protocols, so as to realize intelligent identification, location, tracking, monitoring and management of items[2]. The Internet of Things is called the third revolutionary innovation of the world information industry after the computer, Internet and mobile communication networks[3-4]. In short, the Internet of Things is "a network of things connected". The McKinsey Research Report pointed out that the Internet of Things will be one of the future disruptive technologies, and the Internet of Things has a good development prospect due to its wide application[5]. Driven by innovative technologies, the business model (wearable devices, smart retail, and the sharing economy) with Internet of Things as the main tool is replacing the industrial Internet of Things as the mainstream of IoT commercialization, and even becoming the new force that leads the rapid development of the Internet of Things in the next few years. At the World Economic Forum Annual Meeting 2015, Google boldly predicted that the Internet is about to disappear, and a highly personalized, interactive and interesting world (Internet of Things) is about to be born. According to the statistics of Internet Data Center, the total investment in the global Internet of Things market will reach 1289.9 billion US dollars by 2020. In the next few years, the Internet of Things will expand rapidly worldwide, play an important role in

promoting global economic growth, and may even become the driving force for the fourth industrial revolution[6]. In China, the application and market scale of the Internet of Things industry in transportation, medical treatment, electricity and other fields are increasing year by year, while the level of the Internet of Things industry in China is still low by compared with European and American countries[7]. in the wave of the Internet of Things sweeping the world, competition in technology and market has followed. Keeping pace with the frontier of the IoT research and development, identifying the potential technological opportunities of the IoT is of great significance for China to occupy the core position of future information technology R&D and enhance international competitiveness.

Section 2 reviews the literature about technology opportunity. Section 3 provides details on our data sources and research setting. Analysis from the perspective of technical and market are respectively discussed in detail in Section 4 and Section 5. Section 6 presents concluding remarks.

2 Literature review

At present, global technology competition can be summarized as discovering and exploiting technology opportunities before competitors[8]. From the technical point of view, technology opportunities represent the possibility of technological integration and technological innovation to produce new technological forms. From the industrial point of view, technology opportunities determine the direction of technological development and affect the intensity of industrial R&D[9]. Therefore, the ability to identify technology opportunities is considered to be the key for countries and enterprises to occupy the core of R&D competitiveness. With the proliferation of technology and products and the shortening of the life cycle, the ability to discover technology opportunities plays a more important role in the sustainable development of enterprises[10].

Technology opportunities, which was initially known as technology innovation opportunities, refer to important technologies that bring opportunities to industry innovation or change industry research content[11]. Klevorick holds that technology opportunity is the possibility and potential of technological progress in general or specific fields[12]. In 1995, Porter and Detampel formally put forward the concept of technological opportunity, and pointed out that TOA (Technology Opportunities Analysis) is to infer the forthcoming technological form and development trend through the development trend and interrelationship of existing technologies in the field[13]. They combine technology monitoring with bibliometrics to study technology opportunities for emerging technologies. Due to the limitations of bibliometric analysis, TOA uses semi-quantitative and expert opinions to correct the results of bibliometric analysis. The early technology opportunities were determined by expert judgement and entrepreneurial recognition[14-15]. However, as the increased of the complexity of the technical environment, the reliability of this approach gradually decreased. In order to identify technology opportunities, simple bibliometric indicators including the number of publications and citations of technical articles such as journal articles and patents have begun to be applied to relevant research[16]. With the deepening of research, simple bibliometric indicators are too single to identify technical opportunities. Therefore, some studies combine bibliometrics with other methods, including text mining[17-18], semantic analysis[19], and novelty detection techniques[20]. From the perspective of bibliometrics, a widely recognized standard for identifying technology opportunities is proposed by Naoki Shibata, who holds that terms that exist in scientific papers in active research fields but do not exist in patents can be considered as technology

opportunities[21]. Liu J H pointed out that some concepts that have appeared in science and can be applied in the technical system can be found by analyzing the scientific concept sets and its corresponding technical concept sets, so as to discover new technological opportunities[22]. Based on the research of Naoki Shibata and the data of patents and scientific papers, Huang L C analyzed the technology opportunities in the field of solar cells by used text mining method[23]. Wang K identified four technological opportunities of 3D printing technology based on Naoki Shibata's technology opportunity recognition theory[24]. Domestic scholars identified the core elements of product technology by extracting keywords from patent data[25]. In addition, scholars have also studied technology opportunities from the perspective of enterprises and markets. Lee Y build an expert-based technological attribute-application table, and identified the available technology opportunities in SMEs' existing technologies by using two-stage patent analysis method and multiple keyword matching[26]. Chen Y explored the technology opportunities using TDS theory, combined with patent data from three aspects, including technology, competitors and potential markets[27]. Zhang Y believes that the segmentation of technology opportunities in market innovation activities will help to grasp the technology opportunities[28].

Through the review of the concept, development status and research methods of technology opportunities, we find that the early studies on technology opportunities is mainly focus on expert-based qualitative analysis, semi-quantitative analysis and bibliometrics analysis. However, most of the previous studies on technology opportunity focused on the analysis of technology itself, rarely touching on factors beyond technology. In fact, technology opportunity refer to important technologies that bring opportunities to industry innovation or change industry research content. In addition to technology, studies on technology opportunity should put eyes on market areas. Based on this, this paper uses the quantitative analysis method of bibliometrics to study the technology opportunities from both the technology and market perspectives.

3 Data sources and data processing

3.1 research method

This paper studies the technology opportunities of the Internet of Things from both technology and market perspectives.

From the point of view of technology factors, the analysis of technology opportunities is actually the process of identifying and discovering the technology forms and technology development points that have bright prospects in a technological field. Burst word is an intuitive representation of hot issues, which has the characteristics of rapid growth or sudden increase in frequency of use in a short period of time[31]. Lu W H pointed out that the correct identification of scientific and technological burst words is of great significance to the prediction of scientific research trends, research hotspot, technology opportunities, and scientific and technological monitoring[32]. Burst word can be divided into increasing type and weakening type. And the object of technology opportunity analysis is increasing type of burst words[33]. Wang G F used the Burst Word Detection technology provided by Citespace to comprehensively analyzed the burst words in each period based on the scientific paper data, and identified the frontiers of technology research in the field of bio-nuclear magnetics[34]. Based on Naoki Shibata's theory of technology opportunity recognition, this paper regard the technology represented by burst words which appear in scientific papers but do not appear in patent data as the frontier hotspot of future technology. And we calls it

frontier technology opportunity, that is, the new technology innovation opportunity brought by the frontier of scientific research. For the burst words that exist simultaneously in scientific papers and patent data, if the number of scientific paper on this technology continues to grow and the citation rate continues to increase, it indicates that this technology has new needs in technology application or new discoveries in scientific research. This kind of technology is called mid-end technology opportunity, that is, the technological innovation opportunities brought by the combination of existing technology with significant development potential and latest scientific research findings. This paper follows Naoki Shibata's theory, combined with the burst word detection method and the CiteSpace software to identify the burst words in the scientific papers and patents data of the Internet of Things. And then, identify the frontier and mid-end technology opportunities of IoT technology from the new scientific finding and rapid growth of scientific study.

From the point of view of market factors, the identification of technology opportunities is to serve the development of technology, and then to serve the commercialization and marketization of technology. Therefore, from the perspective of industrialization, technology opportunity analysis should include the identification of market opportunities and potential markets for specific technologies fields. Patent family refers to a group of patent documents which have been applied, published or approved by patent organizations in different countries or regions on the basis of the same priority document. And the patent documents in Patent family have the same or substantially the same content[35]. Chang Y H pointed out that the layout of the patent family reflects the market opportunities to a certain extent. The number of patent family represents the importance of the technology and its future market, and the patent family layout can predict the potential market[36]. On the one hand, the number of patents in patent family of given technology is directly proportional to its importance. That is to say, the more the number of patents in the same family, the greater the technical value and market value[37]. Therefore, if the number of patents in patent family continues to increase, it indicates that this technology is widely recognized by the market, and has broad market prospects. On the other hand, the overseas layout of a national patent family often has a certain market orientation, which to some extent represents the potential market situation in this field. The potential market for technology attracts other countries or regions to carry out patent distribution of related technologies locally. Conversely, the number of overseas patents received by a country or region reflects the market potential of the country or region and its market attraction to related technologies. Zhang L L analyzed the global market dynamics and market expansion of Nano Enabled Drug Delivery technology by using the changes in the number of patents in patent family[38]. Based on the analysis of patent overseas layout, Ma T T studied the potential market in the field of dye sensitized solar cell[30]. In summary, the change in the number of patent family of specific technology represents the market prospect of this technology, and the overseas layout of a national patent family represents the potential market for technology. Frontier technology is the new discovery in scientific research, whoes patent layout is still very vague. Thus, this paper only considers mid-end technology in the analysis of market opportunities for IoT technology. In short, we analyze the market opportunities based on the trend of global patent family layout of mid-end technology, and identify the potential market of IoT technology through the patent layout of IoT technologies in various countries.

3.2 Data collection and processing

The patent data and scientific papers of Internet of Things in this paper are respectively derived from Derwent Innovations Index (DII) and core collection of Web of Science. Published by

the American Institute of Science and Technology Information, Web of Science includes more than 12,000 high-quality journals which reviewed by peer expert. As the world's largest international patent information database, DII has been collected tremendous amount patent data covering the chemical, engineering electrical and electronic categories since 1963， which is an important data source for patent measurement analysis. This paper collects scientific papers and patent data of Internet of Things technology by using "TI=(Internet of thing*) or TS= (Internet of thing*)" as search strategy. The time span is set to 2010-2018 and the retrieval time is August 2018. After refining and cleaning the search results, and eliminating incomplete information or unrelated records, 19 059 patents and 7 032 scientific papers are obtained.

4 Technology opportunities analysis of IoT based on technology factors

4.1 Identifying frontier technology opportunities

The Burst Detection function provided by CiteSpace software can clearly calculate the burst words in specific technical fields, and can display the changes of the burst words at the time level with time gradien[39]. At first, importing scientific paper data and patent data into CiteSpace software respectively. And then, after data conversion, noun phrase extraction and burst words detection, the burst words network maps of IoT technical fields in scientific papers and patents perspective are drawn respectively. Finally, adjusting the network maps layout with timeline view (see Fig. 1 and Fig. 2). The font size in network maps reflects the change in frequency of the burst words.

Analysis of Fig. 1. In 2010, the burst words are "smart object", "sensor network", "future Internet", etc., which indicates that the future application of the Internet is explored through the concept of sensor network technology and smart subject. The burst words in 2011 are “wireless sensor network”, “cloud computing”, “communication technology”, “mobile device”, “new technology”, etc, indicates that the basic technology has become a research hotspot. In addition, the concept of "cloud computing" showing researchers' forward-looking for IoT technology. In 2012, the words "energy/power consumption", "big data", "iot application", "iot environment", "daily life", etc., which indicates that Internet of Things technology began to be applied to daily life. However, the energy supply, application scope and application environment of IoT carriers are still problems to be solved urgently. In addition, researchers combine the emerging concept of big data with the IoT, which reflects the trend of intelligent development of IoT. In 2013, the burst words are "iot devices", "smart cities", "energy efficiency", "iot system", "smart devices", "end user", which indicates that researchers are more deeply in exploring the hardware device and terminal energy consumption optimization of the Internet of Things. In addition, smart city puts higher requirements on the IoT system, which makes “IoT system” a new hotspot in academic research. The burst words "iot services", "things application", "smart sity" and "energy harvesting" appear in 2014 and 2015, which indicates that the functional services of the Iot have become new demands at this stage, so that the application of IoT technology in various fields, especially in smart cities, has become the focus of scientific research. What’s more, the energy problem of the IoT carrier has begun to explore the direction of sustainable, long-lasting and more environmental from internal supply to external collection. In 2016, the burst words changed to “wearable devices” and “new challenge”, indicates

that the application of IoT technology is not limited to the large devices of public services, while details related to users have begun to be the focus. And because of the complexity and precision of IoT carriers such as wearable devices, new challenges are coming. in 2017 and 2018, the burst words are "industrial internet", "fog computing", "proposed system", "smart homes", "cellular network", "edge computing" and "artificial intelligence", which indicated that the deep learning technology generated by the development of the times has become a powerful driving force for the development of the Internet of Things to the next stage. At the same time, the layout of smart cities has begun to shift to a more personalized smart home field. Generally speaking, Judging from the scientific research situation in recent years, the IoT technology appear a trend of intelligence and mass. With the rise of deep learning technology, edge computing, artificial intelligence and virtual reality will become a new technology opportunities.

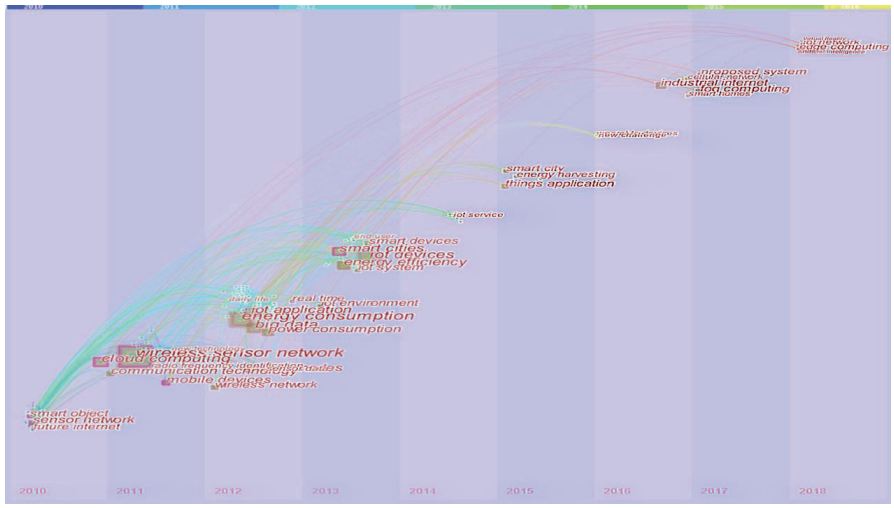


Fig. 1 Bursts words detection result of IoT based on scientific papers

Analysis of Fig. 2. In 2010 and 2011, the burst words are “main body”, “communication module”, “power supply module”, “wireless network”, “GPS”, “RFID”, “input/output end”, “wireless communication”, “mobile phone”, “collection module”, “control device”, “monitoring system”, etc. The patents in this period are mainly about the basic technology of the IoT, indicating that the focus of patent inventors is concentrated on the basic technologies and infrastructure of the IoT. In 2012 and 2013, the burst words are “utility model”, “display screen”, “remote control”, “wireless signal" and “data transmission", which indicates that the patented inventors turned their attention to the IoT device. And the problems of remote control and data transmission were also being solved. In 2014, the burst words are “iot device”, “new utility model”, “real-time monitoring”, etc., which indicates that the IoT device were still the focus of invention. And real-time monitoring technology was applied to IoT devices. In 2015, the burst words are “cloud server”, “electronic device”, “single chip machine”, “intelligent mobile phone”, “wireless connection”, etc. Patent developers have begun to integrate IoT technology with cloud services and implement them through smart devices such as smart phones, which indicates that the research and development hotspots of the IoT have begun to shift to mobile devices. The burst words are “user equipment”, “storage

battery", "of-thing device" and "communication unit" in 2016, which indicates that the research and development of user-based IoT devices has received further attention. In other words, the service of IoT technology to individuals has begun to be realized. In 2017, the burst words are "connecting rod", "box body", "rotating shaft" and "solar panel", which indicates that the refined IoT device has become the main demand of users, so that the IoT device components have become the research focus of inventors. In 2018, the burst words are "programmable logic controller" and "cloud platform", which indicates that the intelligent trend emerging with the combination of IoT technology and new generation of data technology, cloud computing and so on. In summary, at the level of patent R&D, there is a obvious trend that IoT technology has become closer to users in recent years.

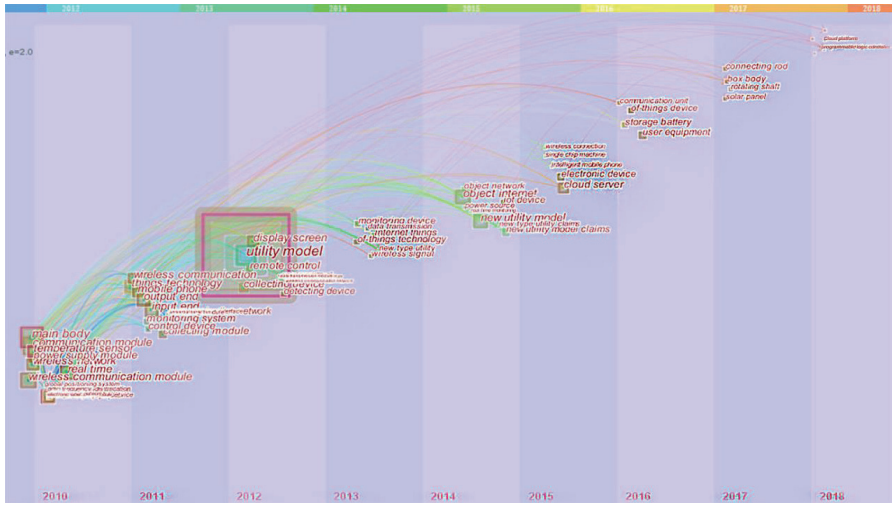


Fig.2 Burst words detection result of IoT based on patent data

Comparative analysis of burst words (see Table 1). In the scientific papers, burst word "cloud computing" first appeared in 2011, while the burst words "cloud server" and "cloud platform" appeared in 2015 and 2018 respectively. In scientific papers, burst word "mobile device" first in 2011, while the burst words "electronic device", "intelligent mobile phone" and "wireless connection" appeared in 2015. In scientific papers, burst word "energy/power consumption" first in 2012, while the burst words "electronic device", "intelligent mobile phone" and "storage battery" appeared in 2016. In the scientific papers, burst words "iot devices" and "smart devices" first appeared in 2013, while the burst words "IoT device", "intelligent mobile phone", "user equipment", "box body", "rotating shaft" and "cloud platform" appeared in 2015, 2016 and 201 respectively. In scientific papers, burst word "energy harvesting" first in 2015, while the burst word "solar panel" appeared in 2017. In summary, the burst words appeared in scientific papers often appeared in patent data in a few years, which confirms Naoki Shibata's theory. Therefore, improving the application of IoT technology in the field of wearable devices and smart homes, promoting the development of algorithms such as fog computing and edge computing in IoT, and encouraging the integration of IoT with artificial intelligence and virtual reality will become the frontier technological opportunities.

Table 1 Comparison of the burst words between of patent data and scientific papers

year	burst words of patent	burst scientific papers
2010	main body, communication module, power supply module, wireless network, GPS, RFID, wireless communication, mobile phone, collection module, control device, real time	smart object, sensor network, future Internet
2011	input/output end, things technology, wireless communication, mobile phone, collection module, control device, monitoring system	wireless sensor network, cloud computing, communication technology, mobile device, new technology
2012	utility model, display screen, remote control, collection device, detection device, data transmission module	energy/power consumption, big data, iot application, iot environment, daily life
2013	IoT technology, monitoring device, wireless signal, data transmission	iot devices, smart cities, energy efficiency, iot system, smart devices, end user
2014	IoT device, utility model, object Internet, real time monitoring	Iot service
2015	cloud server, electronic device, single chip machine, intelligent mobile phone, wireless connection	things application , smart sity, energy harvesting
2016	user equipment, storage battery, of-thing device and communication unit	wearable devices, new challenge
2017	connecting rod, box body, rotating shaft, solar panel	industrial internet, fog computing, proposed system, smart homes, cellular network
2018	programmable logic controller, cloud platform	edge computing, artificial intelligence, virtual reality

4.2 Identifying mid-end technology opportunities

There are some burst words appear in patents and papers at the same time, and the number of scientific papers and citations related to these burst words is increasing rapidly. This phenomenon shows that such technologies have great potential and can be combined with new scientific discoveries to solve new challenges and problems. We called such technologies mid-end technology opportunities. The burst words “IoT devices” and “smart devices” appeared in scientific papers in 2013, and “IoT devices”, “electronic devices” and “user equipment” appeared in the patents in 2014 and 2016. This shows that the scientific research and technological inventions of IoT devices have made new discoveries in the continuous exploration. Collecting scientific papers of IoT in core collection of Web of Science by using “TI=(IoT devices or IoT equipment or Internet of thing* devices or Internet of thing* equipment) OR TS=(IoT devices or IoT equipment or Internet of thing* devices or Internet of thing* equipment)” as search strategy. The time span is set to 2010-2018 and the retrieval time is August 2018. After refining and cleaning the search results, and eliminating incomplete information or unrelated records, 3109 scientific papers are obtained. It can be seen from Fig. 3 that in 2010-2017, the number of scientific papers on “IoT equipment” and its citations are increasing rapidly and continuously without downward trend. This shows that research on the issue of IoT device has always been a hotspot for researchers, and it will also be a hotspot for future technology research and development. It is necessary to point out that the data collection in this

paper is up to August 2018, and the data in 2018 is incomplete. Therefore, the decrease in the citation amount in 2018 in Fig. 3 does not represent a decrease of its research heat. In summary, IoT devices can be considered as a mid-end technology opportunity for IoT technology.

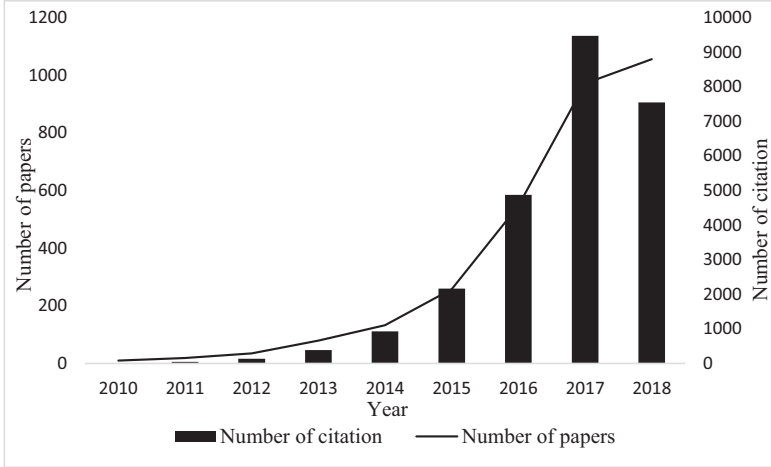


Fig. 3 Changes of the number of scientific papers and its citations in IoT device

The burst words “energy/power consumption”, “energy efficiency”, and “energy harvesting” appeared respectively in scientific papers in 2012, 2013 and 2015, and “power supply”, “storage battery” and “solar panel” appeared in the patents in 2010, 2016 and 2017. This shows that the scientific research and technological inventions of energy issue of IoT devices have made new discoveries in the continuous exploration. Collecting scientific papers of IoT in core collection of Web of Science by using “TS=((internet of thing* or IOT) and (power or energy)) or TI=((internet of thing* or IOT) and (power or energy))” as search strategy. The time span is set to 2010-2018 and the retrieval time is August 2018. After refining and cleaning the search results, and eliminating incomplete information or unrelated records, 1853 scientific papers are obtained. It can be seen from Fig. 4 that in 2010-2017, the number of scientific papers on “IoT energy” and its citations are increasing rapidly and continuously without downward trend. What’s more, it almost doubled since 2014. This shows that research on the issue of IoT energy has always been a hotspot for researchers, and it will also be a hotspot for future technology research and development. Similarly, the decrease in the citation amount in 2018 in Fig. 4 does not represent a decrease of its research heat. Thus, IoT devices can also be considered as a mid-end technology opportunity for IoT technology.

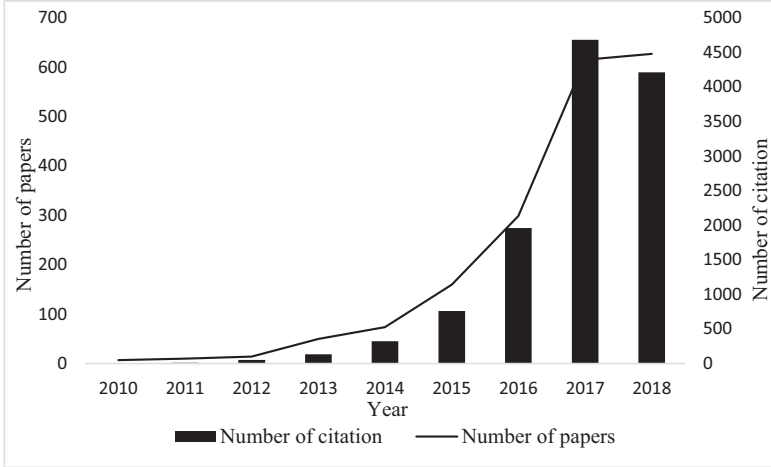


Fig. 4 Changes of the number of scientific papers and its citations in IoT energy

5 Technology opportunities analysis of IoT based on market factors

5.1 Market Opportunity Analysis of Mid-end Technology

Technology is ultimately to serve the market, so it is necessary to prove that mid-end technology has good market opportunities from the market level. Since these two mid-end technologies are technologies that are already in existence and are being studied, their market opportunities could analyzed from the perspective of the layout of the patent family. If the number of patents in patent family is large and rising year by year, this technology is widely recognized in the market, that is to say, it has a good market potential and market prospects. Collecting the patent data of "IoT device" and "IoT energy" in DII from 2010 to 2018 by year stage. And then, processing the collected data to work out the patent family data (see Fig. 5). In 2010-2017, the number of patents in patent family of IoT devices and IoT energy has steadily increased. This shows that these two mid-end technologies have good market prospects and market potential, and can be considered to have good market opportunities.

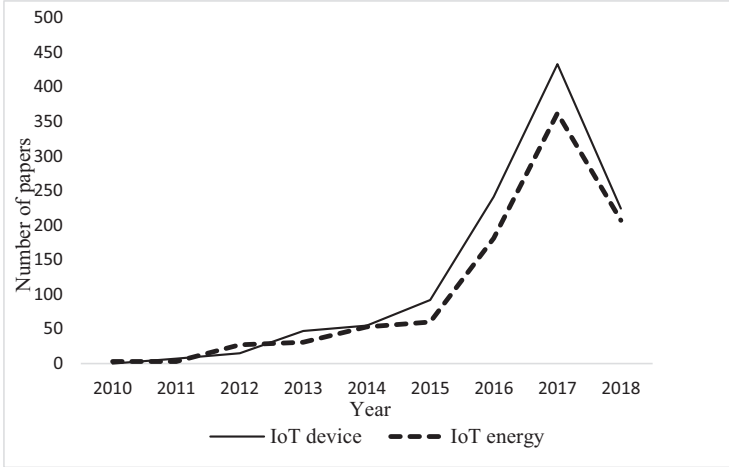


Fig. 5 The trend of the number of patents in patent family of mid-end technology

5.2 Potential market analysis of IoT technology

The overseas layout of patent family reflects the market situation of corresponding technical field. In general, the more patents a country or region distributes overseas, the stronger its technological capability in the corresponding technical fields. The patent overseas layout matrix of IoT technology are constructed based on patent data (see table 2). The number in the matrix represents the patent layout number that row countries distributed in the column countries. It needs to be mentioned that, we are aiming at analyzes the patent layout between countries or regions, so the two organizations such as Would Intellectual Property Organization (WO) and European Patent Office (EP) are not considered. Based on this matrix data, a patent overseas layout network are drawn by using Unicet software (see Fig. 6). Fig. 6 shows that the United States (US), China (CN), and Germany (DE) are located in the center of the network, which indicates that these countries are important IoT technology patent layout countries, and also shows the technical strength of these countries in the field of IoT.

Table 2 patent overseas layout matrix of IoT technology

	CN	US	KR	IN	TW	JP	DE	GB	RD	...	SE
CN	820	25	5	1	3	3	1	0	0	...	0
US	294	1 196	337	120	146	108	9	5	1	...	1
KR	3	22	980	0	0	3	0	0	1	...	0
IN	0	9	0	169	0	0	0	0	0	...	0
TW	0	1	0	0	17	0	0	0	2	...	0
JP	1	1	0	0	0	66	0	0	0	...	0
DE	15	21	6	0	7	2	30	4	0	...	0
GB	1	7	0	0	2	0	1	6	0	...	0

RD	0	0	0	0	0	0	0	0	8	...	0
...	...	...	...	...	...	...	...	...	...	...	...
SE	0	0	0	0	0	0	0	0	0	...	1

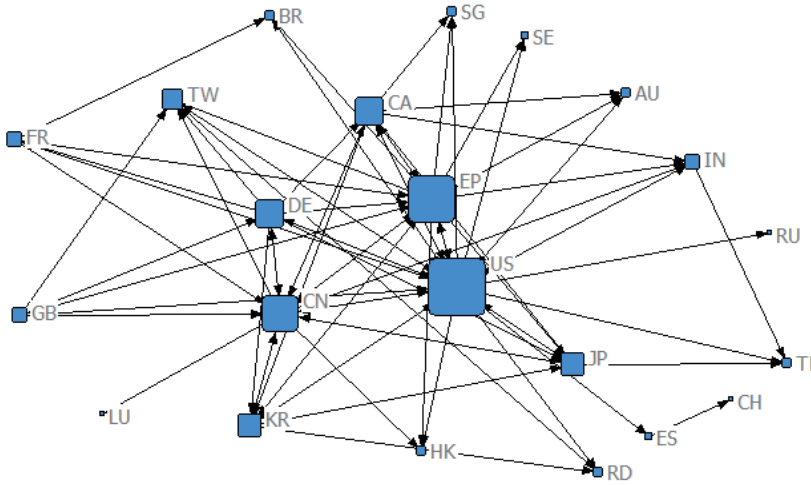


Fig. 6 patent overseas layout network of IoT

The greater the proportion of a country or region receiving overseas patents, the greater the market appeal and market potential for the country. Based on the patent layout matrix data, the patent output rate (PO) and patent receive rate (PR) of each country in IoT technology are calculated according to formula (1) and formula (2). P_i represents the total number of patents IoT in country i, O_{ij} represents the number of patents of IoT of country i distribute in country j, and R_{ij} represents the number of patents of IoT of country i received from country j. Table 3 lists the patent output rate and patent receive rate of countries or regions with more than 20 patent layouts.

$$PO_i = \frac{\sum O_{ij}}{P_i} \tag{1}$$

$$PR_i = \frac{\sum R_{ij}}{P_i} \tag{2}$$

Table 3 Patent output rate (PO) and patent receive rate (PR) of IoT (portion)

	CN	US	KR	IN	TW	JP	DE	GB
PO/%	0.4	70.5	3.3	85.5	3.2	4.1	123.8	61.9
PR/%	2.6	11.0	37.8	79.7	90	89.2	41.9	42.9

Table 3 shows that the PR of China (CN), Korea (KR), Taiwan (TW), Japan (JP) and other countries or regions is higher than their PO, which shows their lower level of internationalization of IoI technology development. Nevertheless, it reflects these countries or regions' strong market attraction and market potential in other view. China has a large number of IoT technology patents, but China's PO and PR are only 0.4% and 2.6%. The small number of PO and PR is due to the huge patent base of IoT technology in China. In fact, the number of overseas patents received by the Chinese market is larger than the total number of patents by some countries, which reflects the great attraction of the Chinese market. The United States (US), India (IN), Germany (DE), and the United Kingdom (GB) have a higher PO than PR, which reflects their status of technology exporting countries. As an important technology exporter in the field of IoT, although its PO is significantly larger than PR, the number of overseas patents it receive is still very large. herefore, the United States is not only an important technology exporter, but also has strong market appeal and market potential. In general, China, the United States, South Korea, Taiwan and Japan have strong technological development momentum and market potential.

6 conclusion

(1) The development of Internet of Things technology has gone through the development of basic technology development and architecture construction, information transmission, IoT carrier/equipment, energy supply, smart devices and cloud services. In this process, scientific research has always led the development of its technology. The latest scientific research on wearable devices, smart homes, fog computing, edge computing, artificial intelligence and virtual reality for IoT technology has become a burst of sudden increase in frequency. According to Naoki Shibata, these can be considered as front-end technology opportunities. IoT device and IoT energy related technologies are simultaneously and continuously explored in scientific research and technology research and development, which indicates that new problems or new research findings have emerged. Thus, these two technologies are considered as mid-end technology opportunities. And these two technologies have good market prospects and market potential.

(2) The United States, China, and Germany are the major technology powers in the Internet of Things. At the same time, the United States, China, South Korea, Taiwan, and Japan have strong market appeal and market potential. Through the analysis of the patent layout of IoT, it is found that the development of the global Internet of Things technology is extremely uneven. The United States, China, and Germany have a large amount of patents. The patent output rate in the United States is relatively high while is extremely low in China. IoT technology in Germany is relatively balanced, with strong technical strength and a relatively stable market. The United States, India, Germany, and the United Kingdom are important exporters of technology. The level of internationalization of IoT technology development in China, Taiwan, Korea and Japan is insufficient, but it has strong market potential. For China, it is necessary to further improve the level of research and development, pay attention to overseas markets with large market potential, and expand overseas layout actively.

Reference

[1] Qian Z H, Wang Y J. IoT Technology and Application[J]. Acta Electronica Sinica, 2012, 40(05): 1023-1029.

[2] Jin J J, Liu J F.Application of New Generation Information Technology in Intellectual Property Service[J].Intellectual Property,2013(06):72-74.

[3] Stankovic J A. Research directions for the internet of things[J].IEEE Internet of Things Journal, 2014, 1(1): 3-9.

[4] Atzori L, Iera A, Morabito G. From “smart objects” to “social objects”, the next evolutionary step of the internet of things[J].IEEE Communications Magazine, 2014, 52(1): 97-105.

[5] Cui Z H, Li H Y. Study on Internet of Things Development Status[J]. Communications Technology, 2014(8):841-846.

[6] Sinha R S, Wei Y, Hwang S H. A survey on LPWA technology: LoRa and NB-IoT[J]. Ict Express, 2017, 3(1):14–21.

[7] Hou J J, Bai Y. Research on the insufficient supply of networking standards: taking the development of key technologies standards as an example[J].Science & Technology Progress and Policy,2015,32(12):61-66.

[8] Yoon J, Park H, Seo W, et al. Technology opportunity discovery (TOD) from existing technologies and products: a function-based TOD framework[J]. Technological Forecasting & Social Change, 2015, 100:153-167.

[9] Olsson O. Technological opportunity and growth[J]. Journal of Economic Growth, 2005, 10(1):31-53.

[10] Feng L, Yu X, Zhang J R.Analysis of High-speed Rail Technology Opportunity Based on Patent Information[J].Journal of Intelligence,2015,34(12):95-100.

[11] Schwartz P. Technological innovation opportunities[J]. Computers and people, 1974, 23 (5):33-33.

[12] Klevorick A K, Levin R C, Nelson R R, et al. On the sources and significance of interindustry differences in technological opportunities[J]. Research Policy, 1995, 24, 185–205.

[13] Porter A L, Detampel M J. Technology opportunities analysis[J]. Technological Forecasting and Social Change,1995,49(3):237255.

[14] Baron R, Ensley M. Opportunity recognition as the detection of meaningful patterns: evidence from comparisons of novice and experienced entrepreneurs[J]. Manag. Sci. 2006, 52, 1331-1344.

[15] Chiu Y J, Chen H C, Tzeng G H, et al. Marketing strategy based on customer behaviour for the LCD-TV[J]. Int. J. Manag. Decis. Mak. 2006, 7, 143-165.

[16] Albert M B, Avery D, Narin F, et al. Direct validation of citation counts as indicators of industrially important patents[J]. Research Policy, 1991, 20(3), 251–259.

[17] Kostoff R. Text mining using database tomography and bibliometrics: a review[J]. Technological Forecasting and Social Change, 2001, 68(3), 223–253.

[18] Gong H Q, Liu Z Q, Huang C.The Research on Discover Robotics Industry Technology Opportunities Based on Patent Text Mining[J].Science & Technology Progress and Policy, 2014, 31(05):70-74.

[19] Yoon J, Kim K. Identifying rapidly evolving technological trends for R&D planning using SAO-based semantic patent networks[J]. Scientometrics, 2001, 88(1), 213–228.

[20] Geum Y, Jeon J, Seol H. Identifying technological opportunities using the novelty detection technique: a case of laser technology in semiconductor manufacturing. Technology Analysis and Strategic Management[J], 2013, 25(1), 1–22.

[21] Shibata N, Kajikawa Y, Sakata I. Extracting the commercialization gap between science and technology-case study of a solar cell[J]. Technological Forecasting and Social Change, 2010, 77(7): 1147-1155.

[22] Liu J H, Jiang Z H. Innovation Drive and Transformation Upgrading under the New Normal[M]. Science

Press, 2016,228-229.

[23] Huang L C, Wang J J, Li X, et al,.Detecting technology opportunities based on papers and patents for perovskite solar cells[J].J Chin Soc Sci Tech Inf,2016,35(07):686-695.

[24] Wang K, Wang J A, Tang Y, et al,.Research on Technology Opportunity Identification Based on Patent and Scientific Papers: An Case of Metal 3D Printing[J].Science and Technology Management Research,2018,38(07):73-79.

[25] Liu J H, Meng Z, Jiang Z H.Research on the Evolution Phases of Toyota Hybrid Electric Vehicle Based on 'Element-Structure-Function-Cost' Framework[J].Science of Science and Management of S.& T.,2017,38(12):26-36.

[26] Lee Y, Kim S Y, Song I, et al. Technology opportunity identification customized to the technological capability of SMEs through two-stage patent analysis[J]. Scientometrics, 2014, 100(1):227-244.

[27] Chen Y, Tan J G, Wang Z Q, et al,. Technological opportunity analysis of industrial robots from the perspective of patents[J]. Science Research Management, 2018, V39(4):144-156.

[28] Zhang Y.Technological Opportunities in the Market Innovation Activities[J].Science and Technology Management Research,2011,31(03):31-34.

[29] Kang Y H, Su J Q.Visualization identify of technology innovation opportunities:empirical analysis based on patent bibliometric[J].Studies in Science of Science,2008(04):695-701.

[30] Ma T T, Wang X F, Zhu D H, et al,.Research on technology opportunity analysis based on patents[J].Studies in Science of Science,2014,32(03):334-342+383.

[31] Chi F, Zhang S Y, Zhao Y, et al,.Burst words detection-based technological opportunities of wearable devices in healthcare[J].Chinese Journal of Medical Library and Information Science,2018,27(03):38-43+64.

[32] Lu W H, Ma J X.Research and Implementation on the Domain Burst Word Recognition Based on CRFs[J].Information Science,2014,32(01):89-93.

[33] Pan D H, Xu K K.Technology opportunity analysis based on co- word analysis[J].Science Research Management,2014,35(04):10-17.

[34] Wang G F, He T, Ma T C, et al,.Biological Magnetic Field Technology Opportunity Recognition Based on Scientific Literature[J].Science and Technology Management Research,2016,36(10):142-147.

[35] Bei F. Interpretation of patent family[J]. Electronics Intellectual Property, 2004(6):52-53.

[36] Chang Y H, Yang W G, Lai K K. Valuable patent or not? Depends on the combination of internal patent family and external citation[C]// Technology Management for Global Economic Growth. IEEE, 2010:1-6.

[37] Wu H, Dong K,Fu X Y.A Comparative Analysis on Patent Intelligence of Fuel Cell in China[J].Journal of Intelligence,2014,33(07):33-37.

[38] Zhang LL, Zhou X, Guo Y, et al,. A Study on Market Expansion Pathways of Emerging Technology Products Based on Patent Data:An Example of Nano Enabled Drug Delivery[J].Journal of Intelligence, 2017, 36(11): 153-160.

[39] Kang C B, Wang W, Zhang S Y, et al,. Research status of studies on medical informatics in China and foreign countries[J]. Chinese Journal of Medical Library and Information Science, 2016, 25(8):24-30.

A New Perspective of Multiple-level for Measuring & Mapping Technology Relatedness

Chunjuan Luan [1] , Bowen Song [2] and Alan L. Porter [3]

[1] *julielcj@163.com*
Faculty of Intellectual Property, Dalian University of Technology, 2# Dagong Road Panjin 124221 (China)

[2] *bowensong333@163.com*
Faculty of Humanities & Social Sciences, Dalian University of Technology, 2# Linggong Road Dalian 116023 (China)

[3] *cluan@dlut.edu.cn*
School of Public Policy, Georgia Institute of Technology, Atlanta, GA 30092, (USA)

Abstract
The extant results are usually limited to reveal technology relatedness from a single IPC, international patent classification code, level. This paper aims at having an insight into technology similarity from a brand new perspective of multiple IPC level, i.e, section level, class level, sub-class level, group level and sub-group level, taking dataset of an emerging technology topic, quantum dot, as an empirical study. Jaccard index is selected to conduct the measurement of technology correlation coefficient for any two IPCs at various levels; social network analysis facilitated by Ucinet software package is employed to draw multiple-level networks. Measurement results show that the more specific the level of IPC is, the higher the mean value is, except section level. Mapping results allow us well informed which IPCs tend to be together; and which IPCs with the highest betweenness centrality values play a significant bridge role in the process of technology converging and developing process. Measuring & mapping technology relatedness from different level can benefit policy-decision makers serving in different level organizations, in aspects such as deploying S&T strategy in macro level, allocating S&T capital and human resources in meso level, and arranging science collaboration and communication in micro level, et al.

Keywords
IPC multiple-level; technology relatedness; correlation coefficient; Jaccard index; measuring and mapping; Girvan and Newman; emerging technology; quantum dot

Introduction
The extant results are usually limited to reveal technology relatedness from a single level (Angue, Ayerbe, & Mitkova, 2014; Luan, Liu, & Wang, 2013). This paper aims at exploring the methodologies and indicators on measuring & mapping technology relatedness for emerging technologies, from a perspective of multiple-level in terms of IPC, i.e., *international patent classification code*. Multiple-level is represented by five IPC level in this study, that is, section, class, sub-class, group and sub-group, the detail procedure of the decomposition of a hierarchical IPC code, an example of H01F-001/01, into 5 levels, is shown in **Figure 1**.

Measuring & mapping technology relatedness from different level, can support policy-decision makers serving in different level organizations, such as national level (Guan & Yan, 2016; Vertova, 1998, 2000), local government level , enterprise level (Angue et al., 2014; Park, Yoon, & Kim, 2012), laboratory level, et al. Emerging technologies are always generated from the extant areas (Daim, Rueda, Martin, & Gerdsri, 2006; Lin, Li, & Wang, 2017; Y. P. Zhang, Lin, & Mi, 2019). From a multiple-level perspective, what is it like the technology relatedness among various fields? Could the accurate similarities among different

areas be worked out? Further, could the relationships be exhibited in a spatial map? Which fields play a significant role in the development of emerging technologies? Co-occurrence analysis (Barbieri & Palma, 2017; Zhao, Wang, & Wang, 2018; Zou, Yue, & Vu, 2018) and Jaccard index (Alkharashi & Evens, 1994; Luan et al., 2013; Y. Zhang et al., 2016) are jointly employed in this study to address the above issues, from 5 IPC levels.

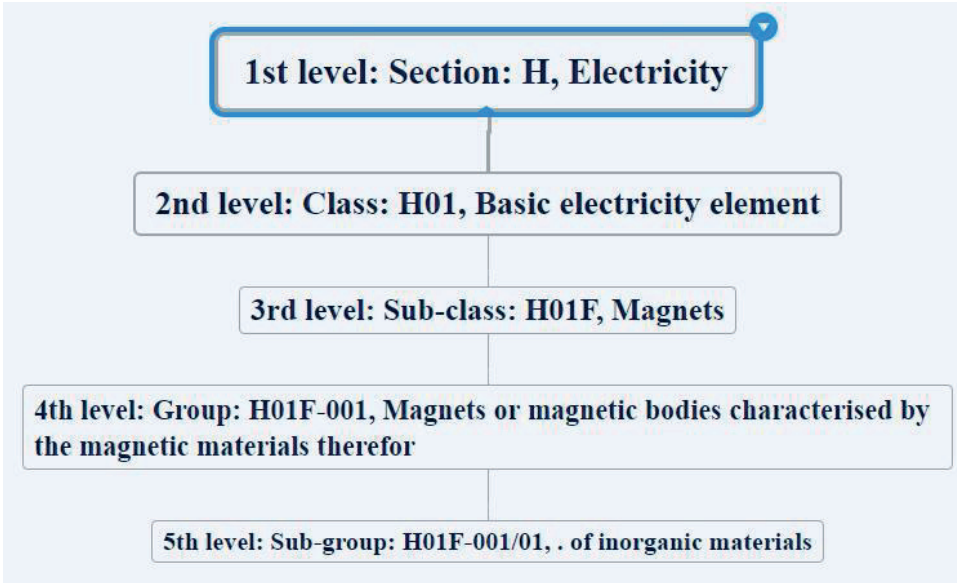


Figure 1. Decomposition of hierarchical IPC code, H01F-001/01, into 5 levels.

As a new generation of nanomaterials (Reshma & Mohanan, 2019; Singh et al., 2018), the theme of quantum dot is selected as the emerging technology in this study. Total 11, 521 patent data are retrieved and downloaded from the database of Derwent Innovations Index on December 26, 2018.

Research approach

Construction multiple-level IPC matrix

Two or more IPC codes are usually indexed in a specific patent document. It is possible to get different level of IPC matrix, by using hypertext software Bibexcel jointly with Ucinet. For example, section-level IPC matrix and sub-group level IPC matrix are demonstrated in **Table 1** and **Table 2**, respectively.

Table 1. Section-level IPC matrix.

	A	*B*	*C*	*D*	*E*	*F*	*G*	*H*
A	0	415	775	16	4	40	477	114
B	415	0	2880	44	14	225	1323	1417
C	775	2880	0	63	18	254	2171	1556
D	16	44	63	0	1	6	16	17
E	4	14	18	1	0	6	17	13
F	40	225	254	6	6	0	620	424

G	477	1323	2171	16	17	620	0	1475
H	114	1417	1556	17	13	424	1475	0

Table 2. Sub-group level IPC matrix.

	B82Y-015/00	*B82Y-020/00*	*B82Y-030/00*	*B82Y-040/00*
C09K-011/65	4	266	166	296
C09K-011/66	3	88	54	91
C09K-011/70	5	102	59	86
C09K-011/88	10	319	222	288

Computation correlation coefficient for multiple-level IPC matrix

Jaccard index is employed to compute correlation coefficient for multiple-level IPC matrix. Leydesdorff (Leydesdorff, 2008) argues that in co-occurrence analysis, Jaccard index should be used. He further interprets that unlike Salton's cosine and the Pearson correlation, the Jaccard index abstracts from the shape of the distributions and focuses only on the intersection and the sum of the two sets. Since the correlations in the co-occurrence matrix may be spurious, this property of the Jaccard index could be considered as an advantage in this case (Luan et al., 2013). According to the Formula provided by Leydesdorff (Leydesdorff, 2008), **Formula 1**) can be used to show the computing method of Jaccard index.

$$S(i,j) = \frac{coo\ (i,j)}{occ(i)+occ(j)-coo(i,j)} \qquad 1)$$

Where S(i, j) represents the correlation coefficient of any 2 IPC i and j, that is, S(i, j) is Jaccard index. Coo(i, j) represents co-occurrence times of i and j; occ(i) and occ(j) represents occurrence frequency of i and j, respectively.

Table 3. Correlation coefficient matrix of Section-level IPC.

	A	*B*	*C*	...	*F*	*G*	*H*	*mean*
A	0	0.024273	0.026022	…	0.003853	0.020932	0.003903	*0.010171*
B	0.024273	0	0.097564	…	0.018689	0.055626	0.047637	*0.031187*
C	0.026022	0.097564	0	…	0.010138	0.060337	0.036481	*0.029254*
…	…	…	…	…	…	…	…	…
F	0.003853	0.018689	0.010138	…	0	0.035638	0.017929	*0.011293*
G	0.020932	0.055626	0.060337	…	0.035638	0	0.041618	*0.02703*
H	0.003903	0.047637	0.036481	…	0.017929	0.041618	0	*0.018617*
mean	*0.010171*	*0.031187*	*0.029254*	…	*0.011293*	*0.02703*	*0.018617*	***0.016285***

Visualization of the multiple-level IPC matrix

Software package of Ucinet and its drawing tool netdraw are applied to visualize the multiple-level IPC matrix in spatial representation. Betweenness centrality networks and main component of the networks are demonstrated level by level, from the 1st level of IPC section, to the 5th level of IPC sub-group.

Multiple-level measurement and visualization

Measuring technology relatedness

According to the research approach delineated in this paper, the accurate value of correlation coefficient between any two IPCs from various levels could be calculated; also, the mean value for any specific IPC in a variety of levels could be worked out in terms of its correlation with the others in the same level. Mean value of correlation coefficient for total 8 sections of IPC is shown in **Table 4**. Mean value of correlation coefficient for top 20 IPCs at the levels of class, sub-class, group and sub-group, are demonstrated in **Table 5** to **Table 8**, respectively.

Table 4. Section-level IPC: mean value of correlation coefficient for total.

Table 5. Class-level IPC: mean value of Correlation coefficient for top 20.

Rank	*Class*	*Correlation coefficient*	*Rank*	*Class*	*Correlation coefficient*
1	F01	0.009665	11	G01	0.003759
2	F02	0.006914	12	H05	0.003594
3	B82	0.006626	13	C12	0.00356
4	F25	0.005615	14	F28	0.003554
5	C09	0.005372	15	G04	0.003493
6	B01	0.004367	16	D02	0.003443
7	G02	0.004212	17	C07	0.003362
8	C01	0.004208	18	G06	0.00313
9	B32	0.004002	19	B29	0.003038
10	B05	0.003922	20	F21	0.00295

Table 6. Sub-class level IPC: mean value of Correlation coefficient for top 20.

Rank	*Sub-class*	*Correlation coefficient*	*Rank*	*Sub-class*	*Correlation coefficient*
1	G02B	0.009329	11	C08K	0.007301
2	C08L	0.009148	12	B05D	0.007124
3	B82Y	0.00914	13	B82B	0.00711
4	F21S	0.009079	14	B32B	0.006718
5	F21V	0.008981	15	B01J	0.006611
6	C08J	0.008847	16	F21Y	0.006609
7	C12Q	0.008807	17	C12M	0.006512
8	C12N	0.007954	18	C08G	0.006269
9	G02F	0.007848	19	C01B	0.0062
10	C09K	0.00739	20	C07H	0.006109

Table 7. Group level IPC: mean value of Correlation coefficient for top 20.

Rank	*Group*	*Correlation coefficient*	*Rank*	*Group*	*Correlation coefficient*
1	A61P-035	0.023742	11	A61K-047	0.015484
2	B82Y-005	0.020156	12	C07K-014	0.015086
3	B82Y-040	0.019193	13	F21V-009	0.014853
4	B82Y-020	0.017821	14	A61K-031	0.014839
5	A61K-049	0.017807	15	A61K-051	0.014827
6	B82Y-030	0.016996	16	A61K-041	0.013476
7	A61K-045	0.016927	17	B82B-003	0.013293

8	A61K-039	0.016848	18	F21S-002	0.013282
9	A61K-038	0.016477	19	G02B-005	0.012996
10	A61K-009	0.016362	20	B82B-001	0.012622

Table 8. Sub-group level IPC: mean value of Correlation coefficient for top 20.

Rank	*Sub-group*	*Correlation coefficient*	*Rank*	*Sub-group*	*Correlation coefficient*
1	C09K-011/08	0.028891	11	G01N-033/53	0.021487
2	B82Y-020/00	0.028015	12	H01L-033/00	0.021011
3	C09K-011/02	0.027616	13	G02B-005/20	0.020986
4	B82Y-040/00	0.026369	14	C09K-011/70	0.019621
5	H01L-033/50	0.025373	15	H05B-033/14	0.019585
6	C09K-011/56	0.023642	16	G01N-033/58	0.019157
7	C09K-011/88	0.023516	17	H01L-029/06	0.018831
8	B82Y-030/00	0.022936	18	H01L-051/00	0.018534
9	H01L-051/50	0.022742	19	G02F-001/1335	0.01798
10	H01L-033/06	0.021522	20	G01N-033/543	0.017488

We further make a comparison of mean "mean value", representing mean value of mean value in Table 4 to Table 8, and obtain mean "mean value" for various IPC level in **Figure 2**.

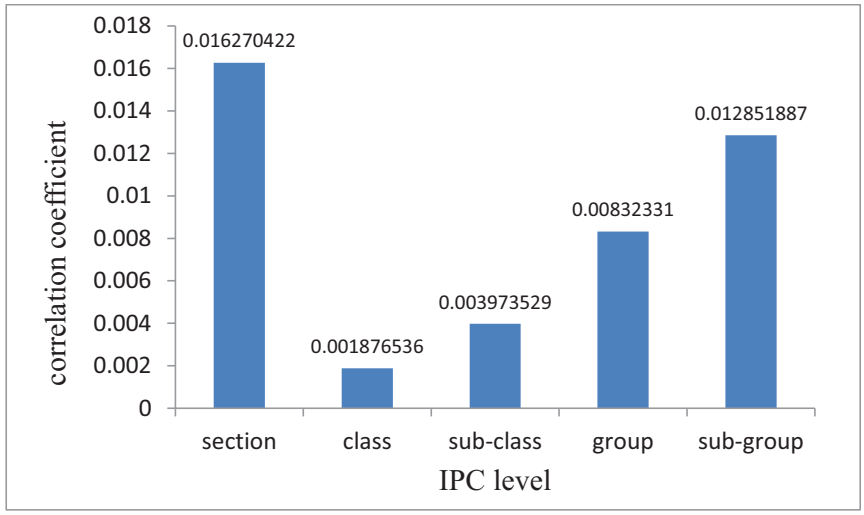


Figure 2. Comparison of mean "mean value" for various IPC level.

Figure 2 reveals a model similar to a V-shaped smile, just a little more to the left. Except IPC section level, the more specific the level of IPC is, the higher the mean value is, from IPC class level, to sub-class level, to group level, to sub-group level. The mean "mean value" for section, 0.016270422, is the indeed mean value for the whole dataset, 11, 521 patent documents. While the other four mean "mean value" are only for top 20 IPCs with the highest mean value in four different levels.

Mapping technology relatedness

Mapping is a spatial representation of the linkage or relationship between different knowledge units or technology fields (Small, 1999). It is focused on monitoring a scientific field and delimiting research areas to determine its cognitive structure or evolution (Noyons, Moed, & Luwel, 1999; Noyons, Moed, & Van Raan, 1999); further to identify the emerging areas (Huang et al., 2017; Pathak et al., 2011; Rotolo, Rafols, Hopkins, & Leydesdorff, 2017). Original co-occurrence matrix data is selected to map for section level IPCs, for only 8 sections totally and the definition is surely guaranteed. For the other four IPC levels, normalized matrix and further (0, 1) matrix is applied for mapping with corresponding threshold to keep it clear enough.

The Girvan–Newman algorithm is used to detect communities in technology network (Girvan and Newman 2002). Girvan and Newman highlighted a property that is found in many networks, the property of community structure, in which network nodes were joined together in tightly knit groups, between which there were only looser connections. They proposed a method for detecting such communities, built around the idea of using centrality indices to find community boundaries. They tested their method on computer-generated and real-world graphs whose community structure was already known, and found that the method detects this known structure with high sensitivity and reliability. Thus, five networks are shown in **Figure 3** to **Figure 7**. The size of the nodes represents the value of betweenness centrality, big or small.

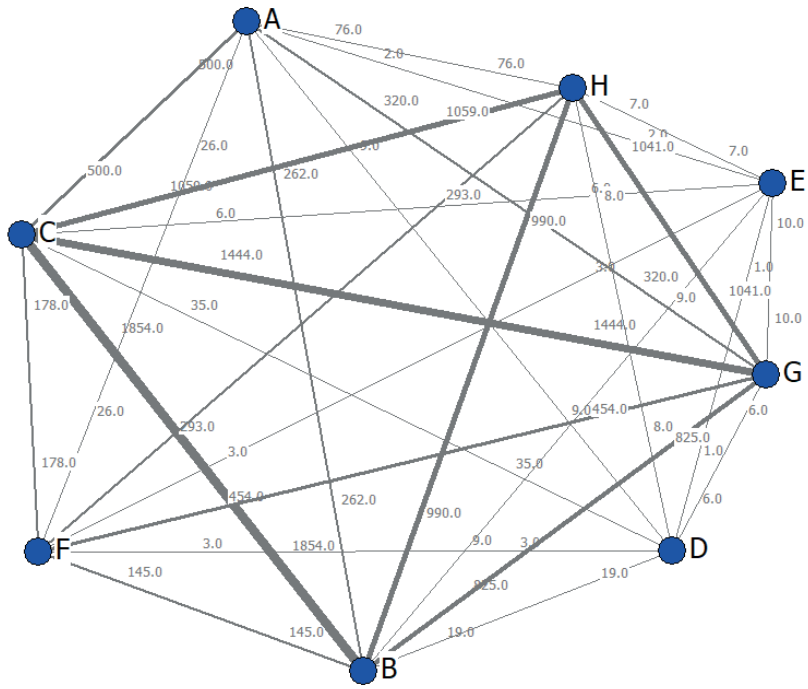


Figure 3. Network of IPCs in section level.

The linkage strength between any two nodes in **Figure 3** is very clear at a glance. The edge weight marked along with the corresponding link reveals that there exists the strongest association strength between Section B and Section C, with strength value 1854; stronger

between Section C and Section G; then Section C and Section H; and then Section G and Section H.

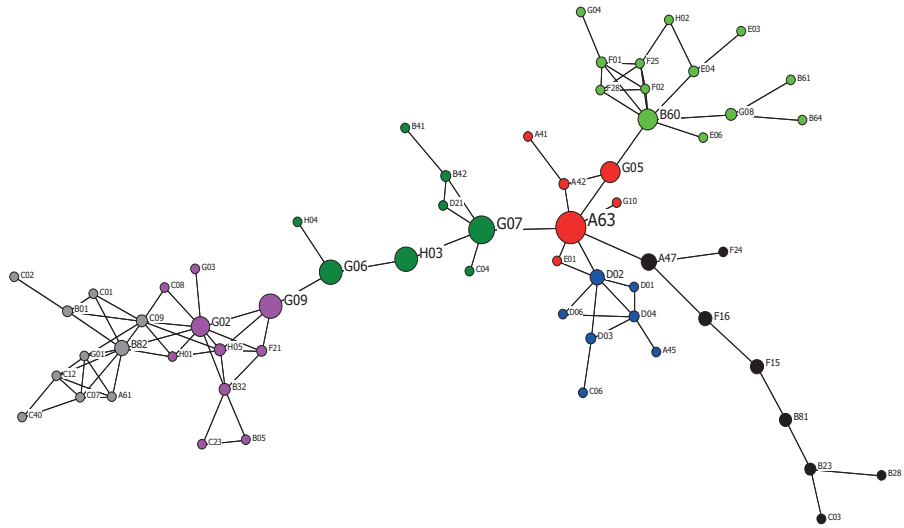


Figure 4. Network of IPCs in class level, threshold=0.02.

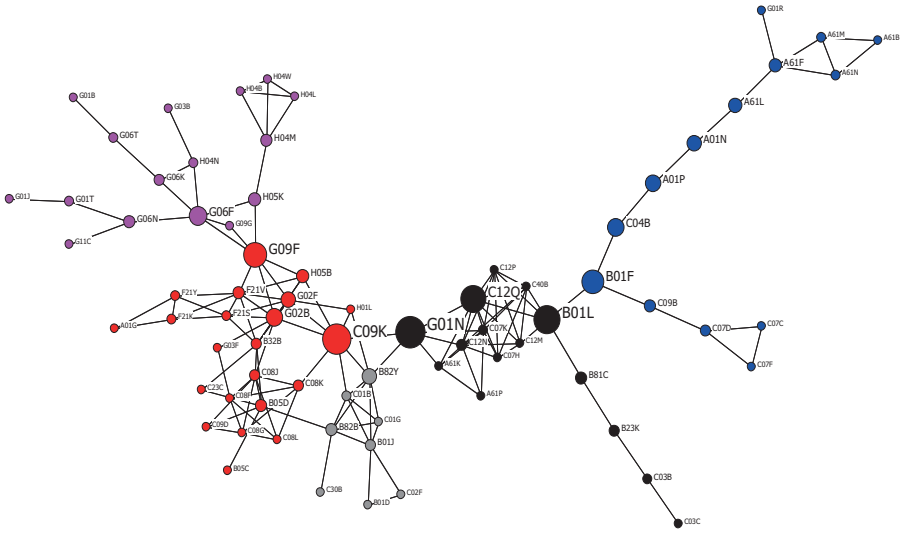


Figure 5. Network of IPCs in sub-class level, threshold=0.025.

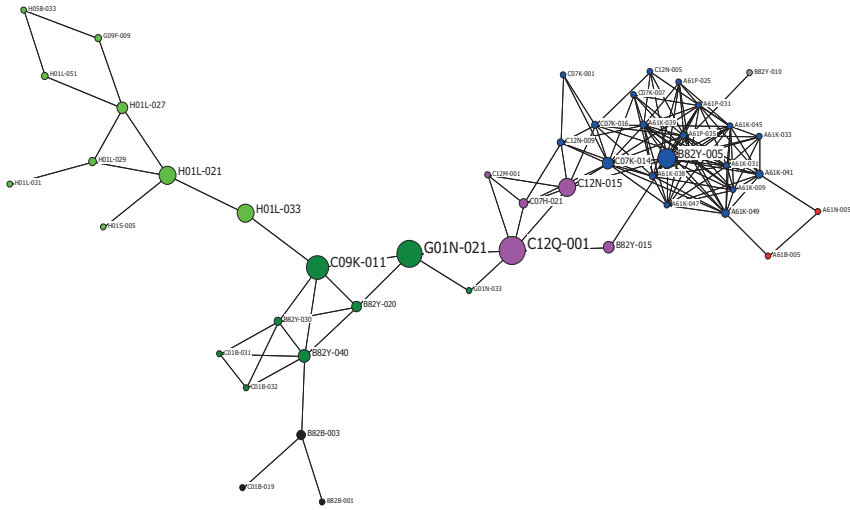


Figure 6. Network of IPCs in group level, threshold=0.05.

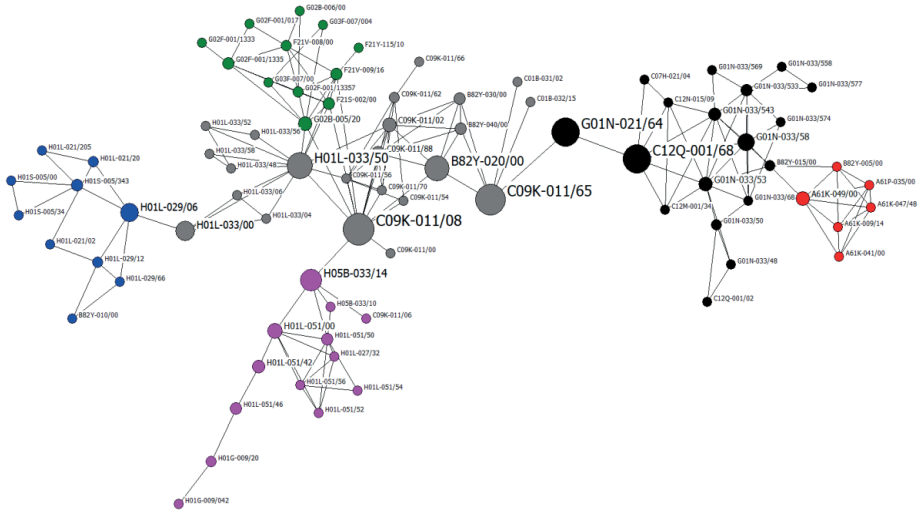


Figure 7. Network of IPCs in sub-group level, threshold=0.08.

Networks from **Figure 4** to **Figure 7** allow us well understand the following information. First, different sub-networks tell us which IPCs tend to be together with comparatively closer connections, having more interactions and usually centred with a specific topic. Second, significant roles the biggest nodes play with the highest betweenness centrality value. Such IPCs bridge various sub-networks, boosting the convergence among different topics or fields.

Conclusion and possible applications

Conclusion

The extant results exploring technology relatedness are usually limited in a single IPC level. In this study, 5-level IPC, i.e. section level, class level, sub-class level, group level and sub-group level, are selected to measure and map technology relatedness; patent publications from an emerging technology field with topic term "quantum dot" are chosen as the dataset. The indicator of correlation coefficient with Jaccard index algorithm is employed to conduct the measurement work, which allow us understand the accurate co-relationship between any two IPCs from various levels. Further, the spatial exhibition of such association strength among IPCs in various levels is visualized by employing Ucinet and Netdraw software, and betweenness centrality jointly with Girvan–Newman algorithm are selected to demonstrate the network.

Measurement results show that the more specific the level of IPC is, the higher the mean value is, from IPC class level, to sub-class level, to group level, to sub-group level; it seems existing a model similar to a V-shaped smile, just a little more to the left.

Mapping results allow us well informed which IPCs tend to be together with comparatively closer connections, having more interactions and usually centred with a specific topic; and which IPCs with the highest betweenness centrality values play a significant bridge role in the process of technology convergence and evolution.

Possible extensions and applications

Measuring & mapping technology relatedness from different level can benefit policy-decision makers serving in different level organizations in different ways. For example, the results of measurement and visualization in IPC section level, being of comparatively macro-level, can be applied for managers serving in topper level organizations, such as National Science Foundation, or central government, to deploy science and technology strategy and improve technology convergence.

Findings of analysis at class level and sub-class level could be used for managers serving in meso-level agencies, such as local government, or university administration, or enterprises. Decision makers could allocate capital and human resources by considering which technology areas playing more significant roles in the corresponding network are worth funding more.

Results of analysis at group level and sub-group level are of microscopic level, such finding could be used in the construction of laboratory in terms of equipment purchasing, employment of human resources from related major, and even arrangement of science collaboration and communication.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (NSFC) under Grant 71473028/71774020/71603040.

References

Alkharashi, I. A., & Evens, M. W. (1994). COMPARING WORDS, STEMS, AND ROOTS AS INDEX TERMS IN AN ARABIC INFORMATION-RETRIEVAL SYSTEM. *Journal of the American Society for Information Science, 45*(8), 548-560. doi: 10.1002/(sici)1097-4571(199409)45:8<548::aid-asi3>3.0.co;2-x

Angue, K., Ayerbe, C., & Mitkova, L. (2014). A method using two dimensions of the patent classification for measuring the technological proximity: an application in identifying a potential R&D partner in biotechnology. *Journal of Technology Transfer, 39*(5), 716-747. doi: 10.1007/s10961-013-9325-8

Barbieri, N., & Palma, A. (2017). Mapping energy-efficient technological advances in home appliances. *Energy Efficiency, 10*(3), 693-716. doi: 10.1007/s12053-016-9470-7

Daim, T. U., Rueda, G., Martin, H., & Gerdsri, P. (2006). Forecasting emerging technologies: Use of bibliometrics and patent analysis. *Technological Forecasting and Social Change, 73*(8), 981-1012. doi: 10.1016/j.techfore.2006.04.004

Guan, J. C., & Yan, Y. (2016). Technological proximity and recombinative innovation in the alternative energy field. *Research Policy, 45*(7), 1460-1473. doi: 10.1016/j.respol.2016.05.002

Huang, Y., Zhu, D. H., Lv, Q., Porter, A. L., Robinson, D. K. R., & Wang, X. F. (2017). Early insights on the Emerging Sources Citation Index (ESCI): an overlay map-based bibliometric study. *Scientometrics, 111*(3), 2041-2057. doi: 10.1007/s11192-017-2349-3

Leydesdorff, L. (2008). On the normalization and visualization of author co-citation data: Salton's cosine versus the Jaccard index. *Journal of the American Society for Information Science and Technology, 59*(1), 77-85. doi: 10.1002/asi.20732

Lin, X. L., Li, Y. B., & Wang, X. Q. (2017). Social commerce research: Definition, research themes and the trends. *International Journal of Information Management, 37*(3), 190-201. doi: 10.1016/j.ijinfomgt.2016.06.006

Luan, C. J., Liu, Z. Y., & Wang, X. W. (2013). Divergence and convergence: technology-relatedness evolution in solar energy industry. *Scientometrics, 97*(2), 461-475. doi: 10.1007/s11192-013-1057-x

Noyons, E. C. M., Moed, H. F., & Luwel, M. (1999). Combining mapping and citation analysis for evaluative bibliometric purposes: A bibliometric study. *Journal of the American Society for Information Science, 50*(2), 115-131. doi: 10.1002/(sici)1097-4571(1999)50:2<115::aid-asi3>3.0.co;2-j

Noyons, E. C. M., Moed, H. F., & Van Raan, A. F. J. (1999). Integrating research performance analysis and science mapping. *Scientometrics, 46*(3), 591-604. doi: 10.1007/bf02459614

Park, H., Yoon, J., & Kim, K. (2012). Identifying patent infringement using SAO based semantic technological similarities. *Scientometrics, 90*(2), 515-529. doi: 10.1007/s11192-011-0522-7

Pathak, J., Wang, J., Kashyap, S., Basford, M., Li, R. L., Masys, D. R., & Chute, C. G. (2011). Mapping clinical phenotype data elements to standardized metadata repositories and controlled terminologies: the eMERGE Network experience. *Journal of the American Medical Informatics Association, 18*(4), 376-386. doi: 10.1136/amiajnl-2010-000061

Reshma, V. G., & Mohanan, P. V. (2019). Quantum dots: Applications and safety consequences. *Journal of Luminescence, 205*, 287-298. doi: 10.1016/j.jlumin.2018.09.015

Rotolo, D., Rafols, I., Hopkins, M. M., & Leydesdorff, L. (2017). Strategic Intelligence on Emerging Technologies: Scientometric Overlay Mapping. *Journal of the Association for Information Science and Technology, 68*(1), 214-233. doi: 10.1002/asi.23631

Singh, R. D., Shandilya, R., Bhargava, A., Kumar, R., Tiwari, R., Chaudhury, K., . . . Mishra, P. K. (2018). Quantum Dot Based Nano-Biosensors for Detection of Circulating Cell Free miRNAs in Lung Carcinogenesis: From Biology to Clinical Translation. *Frontiers in Genetics, 9*. doi: 10.3389/fgene.2018.00616

Small, H. (1999). Visualizing science by citation mapping. *Journal of the American Society for Information Science, 50*(9), 799-813. doi: 10.1002/(sici)1097-4571(1999)50:9<799::aid-asi9>3.3.co;2-7

Vertova, G. (1998). Technological similarity in national styles of innovation in a historical perspective. *Technology Analysis & Strategic Management, 10*(4), 437-449. doi: 10.1080/09537329808524326

Vertova, G. (2000). Technology similarity in national styles of innovation in a historical perspective (vol 10, pg 437, 1998). *Technology Analysis & Strategic Management, 12*(2), U5-U5.

Zhang, Y., Shang, L. N., Huang, L., Porter, A. L., Zhang, G. Q., Lu, J., & Zhu, D. H. (2016). A hybrid similarity measure method for patent portfolio analysis. *Journal of Informetrics, 10*(4), 1108-1130. doi: 10.1016/j.joi.2016.09.006

Zhang, Y. P., Lin, D., & Mi, Z. F. (2019). Electric fence planning for dockless bike-sharing services. *Journal of Cleaner Production, 206*, 383-393. doi: 10.1016/j.jclepro.2018.09.215

Zhao, X., Wang, S. Q., & Wang, X. Y. (2018). Characteristics and Trends of Research on New Energy Vehicle Reliability Based on the Web of Science. *Sustainability, 10*(10). doi: 10.3390/su10103560

Zou, X., Yue, W. L., & Vu, H. L. (2018). Visualization and analysis of mapping knowledge domain of road safety studies. *Accident Analysis and Prevention, 118*, 131-145. doi: 10.1016/j.aap.2018.06.010

International Collaboration in Africa: A scientometric analysis[1]

Radhamany Sooryamoorthy[1]

[1] *sooryamoorthyr@ukzn.ac.za*
School of Social Sciences, University of KwaZulu-Natal, Durban 4041, South Africa

Abstract

Scientific collaboration a key area in scientometric studies. Scholars have researched into several dimensions of scientific collaboration using scientometric data. Such studies have focussed on both individual and group of countries. The objective of this paper is to analyse international collaboration that exists among scholars in Africa. Drawing on the Web of Science data, covering a total of 561,217 publication records for all 53 countries in Africa for the period between 1945 and 2015 the analysis provides an overall view of international collaboration in Africa. The paper also presents recent trends in international collaboration in Africa. The findings show the international connection of Africa with other specific countries in the world.

Introduction

Africa has always attracted scholars from overseas to conduct scientific research in crucial areas of science. Whether it was during or after the colonial period, collaboration prevailed in Africa. Scholars from other countries, the US and Europe in particular, found Africa a fruitful field for science, original data and scientific experiments that opened the doors to numerous discoveries and advancements in science. Africa received scholars and scientists from a large number of countries. Scientific collaboration that exists in Africa has not been detached from its colonial countries. Many African countries continue to maintain their ties with their previous colonisers like the UK, France, Belgium and Portugal. Ex-colonial countries in subSaharan Africa have stronger ties with their former colonisers than with other countries (Nagtegaal and de Brun, 1994). In central African countries, 35 per cent of collaboration occurred with their past colonial rulers (Boshoff, 2009). This historical legacy in collaboration is carried on in both Francophone and Anglophone groups of nations (Adams et al., 2010; Jonathan et al., 2010). In international collaboration, Arab countries (including Algeria, Egypt, Libya, Morocco and Tunisia) have prominent partners in the US, France and the UK (Elalami et al., 1992). Some countries in North Africa including Algeria, Morocco and Tunisia maintain cooperation with European countries such as France due to their past colonial relations (Nour, 2005). International partners of Algeria, Morocco and Tunisia are mostly France, followed by Europe and the US, and for Egypt it is Saudi Arabia (Landini et al., 2015).

Using the Web of Science (WoS) data for the period of 1945–2015, international partnerships that were involved in the publications of African scholars are examined. The analysis is made around the countries with which Africa had scientific contacts and collaboration. This is done for the two separate time periods of 1945–2015 and 2011–2015. A total of 561,217 scientific publications for 1945–2015 and 308,538 for 2011–2015 that had international partnership was analysed.

[1] Due to space constraints only one table could be included in the paper.

Collaboration During 1945–2015

As shown in Table 1, major international partners of Africa are listed according to the number of publications for 1945–2015. The total number of publications does not tally with this number as a publication might have engaged more than one international partner. There were 561,217 publications for this period. The table not only provides the proportionate representation of international partners in African science but also the count of countries in Africa with whom the international partners had research collaboration.

Africa had contacts with most of the countries in the world. A consideration of the first few international partners of Africa with whom the continent produced the largest number of publications shows that they were mostly from North America and Europe.

An exception to this is the presence of two countries from other continents: Saudi Arabia and Australia. The first 10 countries together produced 50 per cent of all publications in Africa for the selected period. Among these, two (the US and France) had more than 10 per cent each of all publications. They were both involved in a quarter of all publications.

France was the second most common international partner in Africa for the entire period of 1945–2015. This has something to do with the colonial connection of France with African countries. Although the data relates to the period 1945–2015 there were not many publications up until 1970. This means the data mostly covers the period of independent Africa. England and Germany were the other colonial partners of Africa that continue to keep their links alive with the new independent Africa.

All the first 10 countries had associated with most of the countries in Africa. The US and England associated with 51 African countries while Germany had collaborated with fewer (36) countries in Africa. On average, foreign partners associated with 39 African countries for the production of scientific publications. There were only 21 countries that associated with Africa to produce 1 per cent or more of its publications in Africa. India, China (1.5% each) and Brazil (1.1%) were among them.

The data also provides information about regional collaboration. Of those countries that were involved in the production of at least 1 per cent of publications, there were only two African countries. South Africa had 2 per cent while Kenya was involved in

1.16 per cent of all publications. Both South Africa and Kenya worked with 47 other African countries.

The association between countries in Africa and international partners shows which of the international partners are closely or loosely collaborated with which African country. Only the top 10 international partners are analysed. The total publications of international partners included all the publications they had produced with Africa and not all of their publications.

The US collaborated mostly with South Africa and Egypt, partnering with 31 per cent and 16 per cent of all the collaborated publications of the US. There were a few other countries with which the US had scientific contacts that were evident in joint publications. The US partners produced 9 per cent of their publications with their counterparts in Kenya, 5 per cent with Nigeria, 4 per cent with Uganda, 3 per cent each with Morocco and Tanzania and about 2 per cent each with Ethiopia, Ghana and Malawi.

France was keen to associate with Tunisia (20% of its publications in Africa), Algeria, Morocco (18% each), South Africa (10%), Cameroon, Egypt, Senegal (4% each) and Burkina Faso (2%). These countries were the French colonies in Africa. Except South Africa with whom France had 10 per cent of its publications with African countries, France did not have serious scientific collaboration with other African countries other than its former colonies. There were also other colonies of France including Gabon, Chad, Niger and Madagascar that did not have much joint scientific production with France. These are not scientifically strong countries in Africa. In other words, scientific association with France is not influenced by the colonial legacy of countries alone

but also by their standing in science. The principal collaborator of England in Africa was South Africa with which it produced one-third of all its publications with African countries. Kenya became the second major partner of England with 10 per cent of publications. Following this were Egypt (9%), Nigeria (6%), Tanzania (5%) and Uganda (4%) These were the British colonies of colonial Africa. Notably England did not have much association with most of the formerly French colonies such as Algeria, Morocco, Senegal and Tunisia. England had significant association which goes back to the period when Egypt was the condominium of both France and England.

Germany's leading scientific collaborators were South Africa and Egypt, producing 32 per cent and 21 per cent of its publications in Africa. South Africa and Egypt were not colonies of Germany. This has to be explored in the second dataset referring to the recent five years. Germany had a long standing relationship with South Africa, mainly during the apartheid period. Kenya, Morocco (5% each), Nigeria (4%), Algeria, Cameroon (3% each), Ethiopia (2.5%), Ghana, Tanzania and Tunisia (2% each) were other countries that Germany favoured for joint science production.

Cameroon and parts of Tanzania were part of Germany at some point in time history.

Saudi Arabia published 75 per cent of its publications in Africa with Egypt alone.

It also published with Tunisia (6%), Algeria (5%), South Africa (4%), Morocco and Sudan (3% each). Saudi Arabia's collaboration with many of these countries is due to its contacts with Arab and Muslim countries.

Although Canada was not a coloniser of Africa, it established scientific contacts with several African countries. Canada produced its publications in Africa mostly with countries such as South Africa (33%), Egypt (16%), Morocco (8%), Kenya (7%), Tunisia (5%) and Uganda (3%). South Africa and Egypt were the two key countries in Africa that advanced in S&T. This might be a cause for Canada to work with scientists from these countries. Similarly, Australia worked with South Africa to produce more than half of its publications with African countries. Guinea (7%), Egypt, Kenya, Morocco (5% each), Nigeria (3%) and Tanzania (2%) were some other countries with which Australian scientists worked.

Italy had a few colonies in Africa (Eritrea, Ethiopia, Libya and Somalia). In joint scientific publications, Italy chose to work mostly with South Africa (producing 26% of its publications in Africa), Egypt (15%), Morocco (11%), Tunisia (9%), Algeria (6%), Nigeria (5%), Kenya (4%) and Cameroon (3%). Seemingly Italy preferred to work with some scientifically strong countries in Africa. The Netherlands produced one-third of its publications in Africa with South Africa, followed by Kenya (9%), Egypt (7%), Morocco, Tanzania (5% each), Ethiopia, Ghana (4% each), Nigeria, Uganda and Zimbabwe (3% each). South Africa had been in touch with the Dutch for a long time.

The Congo, Rwanda and Burundi were colonies of Belgium. The scientific partners of Belgium in Africa among others were South Africa (produced 22% of its publications in Africa), Egypt (10%), Kenya (7%), Algeria, Morocco, Tunisia (5% each), Cameroon, Ethiopia, Tanzania (4% each), Benin, Burkina Faso, Congo, Senegal and Uganda (3% each). Belgium had 2.2 per cent publications with Rwanda and 1.4 per cent with Burundi.

Individually, the colonial legacy, religious or language affinity are not the determining factors that facilitate joint scientific endeavours between countries. The stage at which the countries are in scientific development is a further decisive condition for scientific association and joint publications. As seen in the above data for 1945–2015, former colonial powers while maintaining their links with their former colonies also had scientific contacts with other African countries.

Collaboration During 2011–2015

For this period there were 308,538 publications. Compared to the publication trends revealed in the first dataset for the period of 1945–2015, new patterns are emerging in Africa in scientific

collaboration. Still Africa attracts most of its international partners from North America and Europe. Of the first 10 international partners of Africa there were also countries such as Saudi Arabia, Australia and South Africa. These 10 countries jointly produced 44 per cent of all publications that Africa produced with international partnerships. The US alone had 10 per cent of the publications. In the order of the number of publications were France (8%), England (6%), Saudi Arabia, Germany (4% each), and all other six countries had 2 per cent each. The share of the publications of these countries has been altered during recent years. South Africa is the only country in Africa among the top 10 countries of international partnership.

International partners of Africa collaborated with an average of 31 African countries to produce joint publications. The US worked with 50 countries in Africa while Saudi Arabia associated with 38 countries. More than one per cent of the joint publications with Africa came from 24 foreign countries. Among them were China, India, Brazil and Kenya. South Africa and Kenya were the only African countries that appeared on the list of countries with at least 1 per cent of publications. Since the dataset is for a shorter period of 5 years, countries with at least 0.5 per cent of their publications in Africa should be considered. Half a per cent of publications gives a figure of 1,529 papers. There were 51 countries that had half-a-per cent or more of their publications with African countries. Eight countries were from Africa that represented regional collaboration within the continent. Along with South Africa and Kenya, other African countries were Nigeria, Uganda, Tanzania, Ghana, Cameroon and Morocco. Some Asian countries such as Malaysia, Taiwan, South Korea, Pakistan and Thailand also partnered with Africa, producing between 0.5 and 1 per cent of their publications with Africa. East and West European countries (Russia, Czech Republic, Serbia and Romania) have also taken part in scientific collaboration with Africa. Three South American countries included among these 24 countries were Columbia, Argentina and Chile.

Scientific collaboration between specific countries in Africa and international partners is also examined. All the top 10 international partnering countries are included in the analysis.

In the recent 5-year period (2011–2015), the US produced most of its publications in Africa with a few select countries. From this data, scientific association between African countries and the US can be assumed and explained. South Africa was the major partner of the US, jointly producing 30 per cent of all the publications that the US had produced with African countries. Egypt and Kenya were the second and third research partners, accounting for 15 per cent and 9 per cent. With Uganda the US produced 5 per cent of the latter's publication. The interest of the US in Nigeria was evident in 4.2 per cent of the publications. Tanzania had a similar percentage of publications with the US (3.7%). Other countries that the US favoured for collaboration were Ethiopia, Ghana, Morocco (3% each), Malawi and Zambia (2% each).

France has chosen Tunisia as its major research partner in Africa, producing the highest number of publications in Africa (22%). Collaboration with Algeria led to the production of 17 per cent of publications, 15 per cent with South Africa and 13 per cent with Morocco. Other African countries with whom French associated were Egypt (5%), Cameroon, Senegal (4% each), Reunion (3%), Benin, Burkina Faso, Kenya and Madagascar (2% each).

Producing one-third of all its publications in Africa, England has found its best research partner in South Africa. No other country in Africa had published even half of South Africa's publications with England. England made 10 per cent publications with Egypt, 8 per cent with Kenya, 5 per cent each with Tanzania and Uganda, 4 per cent each with Nigeria and Morocco, and 3 per cent each with Malawi, Ghana and Ethiopia.

Saudi Arabia, in association with Egypt, published three-quarters of all its publications in Africa. Egypt has been the biggest African partner of Saudi Arabia. A few other countries with which Saudi Arabia worked in a significant way were Tunisia (7%), Algeria (5%), South Africa (4%) and Morocco (3%). The rest of the countries in Africa did not have much contact with Saudi Arabia for scientific research.

Germany's key partners in Africa in the order of the percentage of publications were South Africa (31%), Egypt (22%), Morocco (7%), Kenya (5%), Nigeria, Cameroon and Tanzania (3% each). For its publications in Africa, Australia published mostly with South Africa (47%), Egypt, Morocco (7% each) and Kenya (5%). The greatest African partner of Australia was South Africa with which it made nearly half of its publications in Africa. Italy chose to work mostly with South Africa (28%), Egypt (17%), Morocco (11%), Tunisia (10%), Algeria (4%), Cameroon, Kenya and Nigeria (3% each). Spain, which was not among the top 10 countries of international publishers during 1945–2015, worked with countries such as South Africa (27%), Morocco (18%), Egypt (14%), Tunisia (12%) and Algeria (6%). The percentages of publications with other countries were very small.

Canada chose two countries in Africa for most of its joint publications. It worked with South Africa (34%) and Egypt (17%) for its major share of (51%) of publications in Africa. Some other partners of Canada were Morocco (9%), Kenya (6%), Tunisia (5.4%), Uganda (4%) and Algeria (3%).

South Africa is the only African country to appear on the list of the first 10 countries of international partners in Africa. The preferred countries for South Africa were Nigeria (13%), Kenya (11%), Morocco, Zimbabwe (7% each), Uganda (6%), Cameroon, Tanzania (5% each), Ghana, Malawi and Namibia (4% each).

Changes in International Collaboration

There were a total of 561,217 publications during 1945–2015 which involved international partners. For the selected 5-year period of 2011–2015 there were a total of 308,538 publications. That is, 55 per cent of the publications, according to the classification of international partners, were published in the most recent 5 year period. Were there any shifts in international partnership in Africa? Two data sets (1945–2015 and 2011–2015) should be compared for this information.

The percentage of publications is an indicator in this analysis. The first 10 countries can be considered for examination. The first 10 countries with which Africa maintained contacts remained more or less the same in the second period. The only change was that two new countries, namely, Spain and South Africa, made it into the list of top 10. They were not among the first 10 countries during 1945–2015 and moved the Netherlands and Belgium to 11th and 13th positions respectively. During 1945–2015 the US produced 12 per cent of all publications in Africa that had international partnership. This had declined to 10 per cent in the past 5 years (2011–2015). For France, the percentage had been reduced by 3 per cent. England also declined its publication count in the recent 5 years. But for Saudi Arabia there was an improvement from 3 to 4 per cent. No significant change was observed in the percentage of publications of Australia or Italy but Canada lost slightly. The Netherlands and Belgium declined their percentages of publication during 2011–2015. When the first 20 countries were examined, a few more changes were observed. Spain strengthened both its position (from 12th to 8th) and share of publications. Switzerland, Japan, Sweden and Scotland decreased in the percentage while China, India, Brazil and Kenya increased their respective shares of publications during 2011–2015. Have there been any increases or decreases in the average number of countries in Africa with which international collaborators published? The average decreased from 39 to 31 countries recently.

For the US, its African partners with whom most of its publications in Africa were published have not changed. South Africa, Egypt and Kenya remained the first three partners of the US. The US works more closely now with Uganda than Nigeria if the percentage of publications is any indication. The US improved its association with Tanzania, Ethiopia and Ghana by producing slightly more publications in the second period. The key partners of France remained Tunisia, Algeria and Morocco. However, it increased its production with Tunisia and South Africa while it

decreased publications with Morocco. With Cameroon, Egypt and Senegal, France maintained its share of publications in Africa.

In the recent period England was also the third most highly published country with African countries. With South Africa its share of publications did not change between the two periods. In second place of partnership was Kenya and not Egypt as in the first period. With Nigeria its share declined by 2 per cent and with Uganda it increased by 1 per cent.

Saudi Arabia moved up one position, becoming the fourth country with the highest number of publications with Africa. No change occurred in the percentage of publications of Saudi Arabia with Egypt, South Africa, Algeria and Morocco between the two periods. Germany occupied the fifth position. Germany also maintained its scientific association with South Africa, Egypt and Kenya but this slightly declined with Nigeria. It improved its links with Morocco and Tanzania.

South Africa continued to be the best partner of Australia, producing about half of the latter's publications in Africa. The new partners of Australia included Egypt, Morocco (7% each) and Guinea (5%) with a sizable percentage of publications. Italy continued to work mostly with their main partners such as South Africa, Egypt, Tunisia (increased in the recent years), Morocco, Cameroon (at the same level), Algeria, Nigeria, and Kenya (declined slightly).

Canada lagged behind by 3 ranks, from the 6th top international partner of Africa during 1945–2015 to 9th top partner during 2011–2015. Canada increased its share of publications with South Africa, Egypt, Morocco and Uganda slightly and decreased with Kenya.

The paper shows scientific collaboration in Africa for a longer period of time, and presents the trends in recent years. This analysis offers insights into the dynamics of scientific collaboration in African countries that had historically associated with other countries on the continent and elsewhere. The findings also indicates the potential for the countries in Africa advance science in their respective areas of strength.

References

Adams, J., King, C., & Hook, D. (2010). *Global research report, Africa 2010*. Leeds: Evidence Thomson Reuters.

Boshoff, N. (2009). Neo-colonialism and research collaboration in Central Africa *Scientometrics, 81*(2), 413-434.

Elalami, J., Dore, J. C., & Miquel, J. F. (1992). International scientific collaboration in Arab countries. In R. Arvanitis & J. Gaillard (Eds.), *Science indicators for developing countries* (pp. 357–371). Paris: ORSTOM.

Jonathan, A., Christopher, K., & Daniel, H. (2010). *Global research report Africa*. Leeds: Evidence, Thomson Reuters.

Landini, F., Malerba, F., & Mavilia, R. (2015). The structure and dynamics of networks of scientific collaborations in Northern Africa. *Scientometrics, 105*, 1787–1807.

Nagtegaal, L. W., & de Brun, R. E. (1994). The French connection and other neocolonial patterns in the global network of science. *Research Evaluation, 4*(2), 119–127.

Nour, S. S. O. M. (2005). Science and technology development indicators in the Arab region: A comparative study of Arab gulf and Mediterranean countries. *Science, Technology & Society, 10*(2), 249–274.

Table 1 Major partners of Africa in science, 1945–2015

Country	No. of Publications	% of all publications	Mean	S.D	No. of African countries
US	68797	12.26	1348.96	3373.69	51
France	59655	10.63	1193.10	2762.12	50
England	42339	7.54	830.18	1966.75	51
Germany	26569	4.73	542.22	1441.08	36
Saudi Arabia	16553	2.95	376.20	1874.38	44
Canada	14648	2.61	292.96	772.14	50
Australia	13503	2.41	277.55	980.29	49
Italy	13492	2.40	269.84	610.75	50
Netherlands	13264	2.36	276.33	679.85	48
Belgium	12298	2.19	250.98	440.97	49
Switzerland	12261	2.18	245.22	529.38	50
Spain	11736	2.09	234.72	598.48	50
South Africa	11475	2.04	244.15	330.46	47
Japan	10272	1.83	223.30	611.31	46
Sweden	9677	1.72	205.89	459.84	47
India	8590	1.53	186.74	465.3	46
China	8280	1.48	176.17	476.88	47
Scotland	7360	1.31	156.60	391.78	47
Denmark	6876	1.23	149.48	315.24	46
Kenya	6507	1.16	138.45	234.97	47
Brazil	5961	1.06	129.59	328.99	46
Norway	5519	0.98	117.43	305.62	47
Austria	5146	0.92	114.36	305.54	45
Nigeria	4579	0.82	95.40	219.32	48
Poland	4497	0.80	121.54	361.51	37
Portugal	4313	0.77	95.84	239.78	45
Russia	3967	0.71	88.16	295.89	45
Uganda	3697	0.66	85.98	161.88	43
Tanzania	3694	0.66	83.95	147.71	44
Czech Republic	3584	0.64	87.41	240.99	41
Israel	3557	0.63	75.68	264.08	47
Greece	3203	0.57	82.13	223.04	39
Cameroon	3089	0.55	65.72	96.26	47
South Korea	3002	0.53	75.05	245.85	40
Mexico	2968	0.53	64.52	169.15	46
Malaysia	2947	0.53	75.56	145.97	39
Finland	2870	0.51	70.00	178.46	41
New Zealand	2868	0.51	65.18	212.45	44
Hungary	2819	0.50	80.54	238.35	35
Turkey	2802	0.50	65.16	194.04	43
Ghana	2725	0.49	61.93	90.57	44
Taiwan	2668	0.48	65.07	187.02	41
Thailand	2615	0.47	60.81	111.09	43
Pakistan	2608	0.46	65.20	150.89	40
Zimbabwe	2553	0.45	63.83	163.44	40
Argentina	2529	0.45	63.23	199.51	40
Senegal	2485	0.44	51.77	61.63	48
Columbia	2454	0.44	63.18	152.04	39
Chile	2410	0.43	60.25	214.7	40
Romania	2361	0.42	69.44	189.09	34
Morocco	2343	0.42	55.79	138.96	42
Egypt	2302	0.41	52.32	87.06	44
Burkino Faso	2248	0.40	54.83	64.84	41
Ethiopia	2087	0.37	47.43	85.36	44
Iran	2051	0.37	50.02	142.57	41
Ireland	1992	0.35	45.27	114.92	44
Serbia	1982	0.35	60.06	160.5	33
Tunisia	1969	0.35	48.02	116.06	41
Malawi	1915	0.34	46.71	85.76	41
Zambia	1887	0.34	43.88	77.2	43
Cote Ivore	1869	0.33	39.77	47.25	47
Armenia	1849	0.33	80.39	197.23	23
UAE	1719	0.31	47.75	159.34	36

Table 1 Major partners of Africa in science, 1945–2015 (cont.)

Country	No. of Publications	% of all publications	Mean	S.D	No. of African countries
Syria	612	0.11	19.13	33.12	32
Togo	568	0.10	16.71	25.55	34
Baharain	513	0.09	16.03	32.03	32
Luxumberg	477	0.08	11.93	14.56	40
Yemen	473	0.08	16.31	55.18	29
Libya	452	0.08	15.07	47.15	30
Iraq	413	0.07	14.75	30.93	28
Cambodia	406	0.07	10.97	10.47	37
Congo People's Rep	375	0.07	23.44	86.56	16
Central African Republic	350	0.06	11.29	15.89	31
Cuba	347	0.06	10.21	26.45	34
Ecuador	313	0.06	7.83	12.73	40
UAE	1719	0.31	47.75	159.34	36
Wales	1674	0.30	38.05	93.58	44
Slovakia	1601	0.29	43.27	142.22	37
Benin	1581	0.28	37.64	47.26	42
Rep of Georgia	1561	0.28	53.83	144.29	29
Mali	1515	0.27	36.07	47.34	42
Bylaurs	1430	0.25	75.26	172.08	19
Algeria	1404	0.25	36.95	96.99	38
Slovenia	1241	0.22	47.73	145.9	26
Indonesia	1230	0.22	29.29	38.48	42
Botswana	1211	0.22	30.28	71.32	40
Bulgaria	1186	0.21	34.88	101.22	34
Qatar	1181	0.21	39.37	128.08	30
Gambia	1153	0.21	27.45	34.75	42
Swaziland	1134	0.20	45.36	159.82	25
Sudan	1125	0.20	26.16	42.1	43
Mozambique	1107	0.20	27.00	46.19	41
Kuwait	1089	0.19	40.33	154.17	27
Ukraine	1081	0.19	36.03	105.48	30
Niger	1071	0.19	23.28	35.14	46
Namibia	1060	0.19	28.65	110.64	37
Croatia	1034	0.18	33.35	103.24	31
Lebanon	1023	0.18	26.92	57.76	38
Azerbaijan	1018	0.18	46.27	134.44	22
Peru	1007	0.18	24.56	49.86	41
Singapore	1007	0.18	25.18	49.24	40
Bangladesh	926	0.16	25.72	31.77	36
Vietnam	924	0.16	23.69	28.83	39
Guinea	916	0.16	24.76	82.35	37
Jordan	906	0.16	26.65	57.05	34
Gabon	876	0.16	23.05	29.74	38
Rwanda	869	0.15	22.87	32.09	38
Oman	850	0.15	23.61	54.89	36
Philippines	844	0.15	20.59	28.37	37
Estonia	818	0.15	23.37	78.86	35
Madagascar	817	0.15	19.00	25.6	43
Congo	789	0.14	18.35	21.32	43
Cyprus	737	0.13	24.57	86.55	30
North Ireland	715	0.13	23.06	65.86	31
Zaire	708	0.13	18.63	20.65	38
Sri Lanka	705	0.13	19.05	36.95	37
Guinea-Bissau	647	0.12	23.11	86.81	28
Lithuania	637	0.11	24.50	85.77	26
Syria	612	0.11	19.13	33.12	32
Togo	568	0.10	16.71	25.55	34
Baharain	513	0.09	16.03	32.03	32
Luxumberg	477	0.08	11.93	14.56	40
Yemen	473	0.08	16.31	55.18	29
Libya	452	0.08	15.07	47.15	30
Iraq	413	0.07	14.75	30.93	28
Cambodia	406	0.07	10.97	10.47	37
Congo People's Rep	375	0.07	23.44	86.56	16
Central African Republic	350	0.06	11.29	15.89	31

Table 1 Major partners of Africa in science, 1945–2015 (cont.)

Country	No. of Publications	% of all publications	Mean	S.D	No. of African countries
Cuba	347	0.06	10.21	26.45	34
Ecuador	313	0.06	7.83	12.73	40
USSR	156	0.03	6.50	11.09	24
Total for all authors	561217	100.00	102.82	258.09	39.33

Note: Total includes all partners that are not shown in the table.

Mean values of skew distributions in bibliometrics

Ulrich Schmoch

Frauhofer ISI, Breslauer Str. 48, 76139 Karlsruhe, Germany. Email: us@isi.fhg.de

Abstract

Nearly all distributions in bibliometrics are skew. In particular, the distribution of citations of publications by research units is skew. For a ranking of research units, it is recommended to replace the calculation of standard mean values by adjusted mean values where outliers with very high positive citations are cut and those with low or no citations are cancelled. Such an adjusted mean value is oriented on the average activity of a research unit and leads to a more adequate assessment. This approach is based on the concept of the Hirsch-Index. This calculation implies a different ranking of research units and it is important in cases where the distribution of finances depends on bibliometric rankings.

Introduction

Mean values and skew distributions are two major topics in the discourse on bibliometrics. Mean values are used for assessing the citation score of organisations or countries or the citation behaviour in specific fields. For calculating the standardization of citations, the mean value of citations per author/organisation is divided by the mean value of the field of publication. Thus in particular for the assessment of citation performance, mean values are regularly used. The skew distribution is a general observation applying to the publications per author or the citations per author. A general formulation of skew distributions in bibliometrics was formulated by Lotka, and skew distributions can be found in all types of bibliometric analysis. Sometimes, the distributions are extremely skew. Against this background, the question may be asked whether the calculation of standard mean values in bibliometrics leads to meaningful results. The paper suggests an alternative approach to calculate mean values and its implications for rankings.

The discourse on skewness

A fundamental early publication on skewness was published by Lotka (1926). In this contribution, the author suggests a distribution of the publications per author according to the formula $X^n = C/Y$ wherein:

X = number of publications
C, n = Constants depending on the specific field (n ~ 2)
Y = the relative frequency of authors with X publications

This means that, e.g., 10 authors among 100 have only 1 publication, 25 authors have about 4 publications and 1 author has 100 publications. A graph with the papers written and the percentage of authors leads two a quite skew distribution with a small number of authors with very high number of publications and many authors with very few publications.
Various studies were conducted to verify the so-called Lotka's Law, e.g., Murphy (1973), Pao (1986) or Radhakrishnan (1973). In most cases, Lotka's Law could be confirmed, if the examined samples are sufficiently large. Subsequent to Lotka (1926), various other suggestions for skew distributions very made, e.g., Chen & Leimkuhler (1986) discussed Lotka's, Bradford's and Zipf's Law, Simon (1955) suggested alternative functions, the latter were modified by Mandelbrot (1959), as general statistical description was presented by Adamic (2002). Skew distributions were found for many areas beyond publication productivity, e.g., for linguistics (Zipf 1949) or income distribution (Pareto 1935). A good overview is provided by Newman (2005). In any case, skew distributions are a frequent phenomenon in science and in

consequence in bibliometrics. In particular, skewness is characteristic for citation pattern (Seglen 1992).
As skewness is an important phenomenon in bibliometrics, some well known authors have discussed it. De Solla Price (1976) described this topic in terms of cumulative advantage in detail, Narin & Hamilton (1996) emphasise the relevance of highly cited publications or patents within skew distributions, Glänzel & Moed (2005) discuss journal impact measures and state a statistical reliability of the comparison of impact factors despite skew distributions of the citations.
A prominent contribution on the quantification of an individual's scientific research output was made by Hirsch (2005). He suggests a simple measure for the scientific performance defined as the maximum value of h such that the given author/journal has published h papers that have each been cited at least h times. This measure de facto implies that extremely high citation values are neglected as well as very low ones. The implications of the H-Index are illustrated in the next section.

H-Index, skew distributions and mean values

The determination of the H-Index is illustrated by the example of a research unit in the area of the application of graphene in electrical engineering. According to a search in the Web of Science, this unit published in 2015 62 articles in scientific journals with a total of 1887 citations in the period of 2015 to 2017. This leads to an average citation rate of 30.4 (mean value). The maximum citation rate is 209, 7 publications are not cited at all until the end of 2017 (cf. Figure 1).
Using the definition of Hirsch, an index value of 27 is determined illustrated in Figure 1 by a bold line. This definition is based on the rank of the publications according to their citation level. Thus, the publications with more than 27 citations are not considered in more detail and their level has no impact on the index. The same applies to publications with low or no citations. To formulate this in a statistical perspective, the H-Index does not consider high and low outliers and is focused on the standard performance of a unit. This perspective can be qualitatively justified by the observation that even high level institutions often have publications with low or even without citations which document the outcome of intermediary working steps (Schmoch, Beckert & Schaper-Rinkel 2019). Due to the steady pressure to publish even these results are published. In the case of extremely high citation scores, the latter are often not the indication of high performance, but a coincidental effect of conducive circumstances. For instance, a paper may be an early contribution to a broad, long-term discourse and every subsequent publication has to publish this early one. In any case, these extreme values are not representative for the average activity of a unit. Such a reflection will be the background of the concept of Hirsch.
It is possible to transfer this reasoning into simple rules of bibliometric analysis. In the case of the exemplary unit, the 20 percent of the publications with the highest citation are cut to the level of the last publication in this percentile and all publications with less than 6 citations are deleted. This amendment lead to the "adjusted distribution" in Figure 1. The mean value of the adjusted distribution is 27.9, thus it is near the Hirsch-Index.

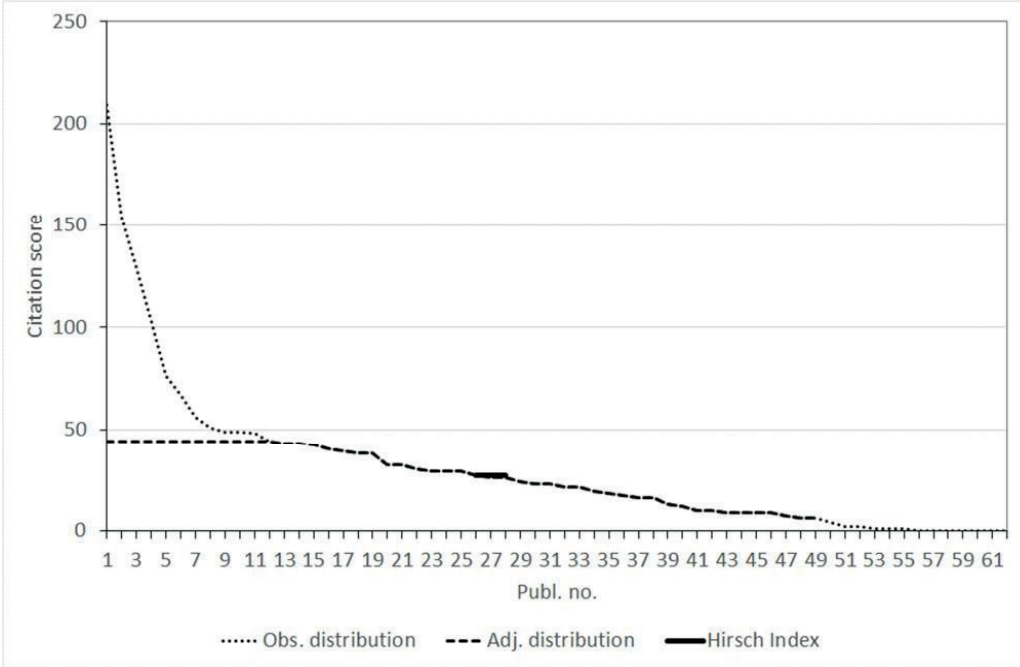


Figure 1. Citations on the publications of a research unit in the area of graphene in electrical engineering in 2015, source: Web of Science, update 2018

All in all, although the calculation of mean values of skew distributions is possible in a mathematical view, there are good reasons to assess the citation performance of research units on the basis of their average activity and to exclude extreme positive and very low values.

Rankings of research units based on adjusted distributions

There are good reasons to calculate the mean value of skew distributions of citations on the basis of adjusted instead of full range distributions. However, the level of resulting mean values is quite similar to the standard mean value, for the example of the former section 27.9 instead of 30.4. Therefore, the question has to be raised, whether this difference is so important that a separate calculation implies new insights and, in particular, whether rankings of research units change. The latter issue is important, as in many countries, the distribution of research funding is based such bibliometric rankings, e.g., for the Research Excellence Framework (REF) in the United Kingdom.

For checking the implication of adjusted mean values on the ranking of research units, we analysed the citation activity of ten research units in the subject category "Biotechnology & Applied Microbiology" of the Web of Science with about 500 citations to publications of the year 2015. In Table 1, these research units are ranked according to the mean value of their citations. The calculation of the adjusted mean values leads to, of course, to different values and to a different ranking, too. Thus, the ranking is generally similar, but, e.g., Unit 3 advances to the second rank, Unit 4 drops to the forth or Unit 8 advances to the sevenths rank. This change in the ranking is due to the fact that the distributions for all research units are skew, but the samples are so skew that degree of skewness differs by unit. To conclude, the calculation of adjusted mean values implies different rankings.

For illustrating the change of ranking, Figure 2 shows the citation distributions for Unit 2 and Unit 3. In the standard calculation of mean values, Unit 2 gets a higher rank due to two publications with very high citations. In the adjusted calculation, Unit 3 achieves a higher position, as it has in the area of standard activities some publications with higher citation scores than Unit 2. Here, one has to decide whether some very high citations or various medium citations in the standard activities are more characteristic to assess a research institution.

Table 1. Ranking of research units in "Biotechnology & Applied Microbiology" with about 500 citations to publications of 2015, source: Web of Science, update 2018

Research Unit	# publications	# citations	mean value	Adjusted mean value	Rank adj. mean
Unit 1	10	506	50,60	63,00	1
Unit 2	22	517	23,50	25,60	4
Unit 3	25	531	21,24	30,00	2
Unit 4	24	507	21,13	26,16	3
Unit 5	28	512	18,29	21,74	5
Unit 6	32	532	16,63	18,61	6
Unit 7	41	516	12,59	17,89	8
Unit 8	42	514	12,24	16,66	9
Unit 9	42	511	12,17	18,00	7
Unit 10	49	515	10,51	15,35	10

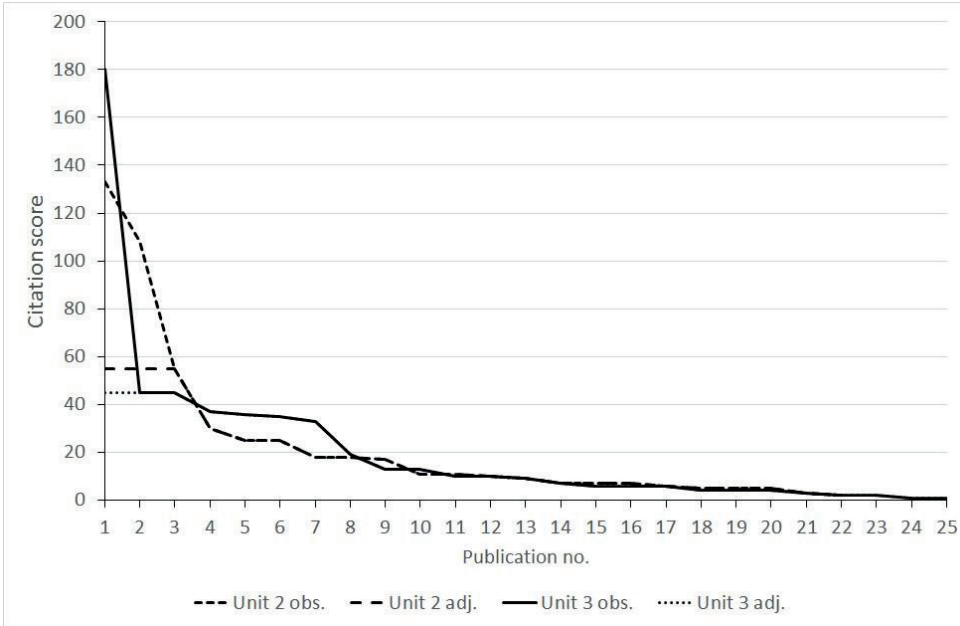


Figure 2. Distribution of citations of two selected research units in "Biotechnology & Applied Microbiology, source: Web of Science, update 2018.

To further illustrate the effect of adjusted means, another example is shown in Figure 3. Here, various research units in physics were compared, in this case in the units with about 300 publications in 2013. From this sample, two units were selected. For highlighting the distribution of the citations, Figure 3 is cut at 30 publications. In the observed distribution, Unit 2 has 313 publications, Unit 6 306 publications. Due to the high number of publications with no citations, the standard mean values of both units are very low and nearly at an equal level. Unit 6 has a little advance due to two high citation scores of 124 and 122. In the adjusted calculation, Unit 2 is above Unit 6 due to the higher citation scores in the standard activities.
This example shows that the adjustment can imply a fairer comparison of research units and that a few highly cited publications should not compensate relevant standard activities. Furthermore, standard mean values for research units with long tail of uncited publications are dominated by these tails and imply significant distortions.

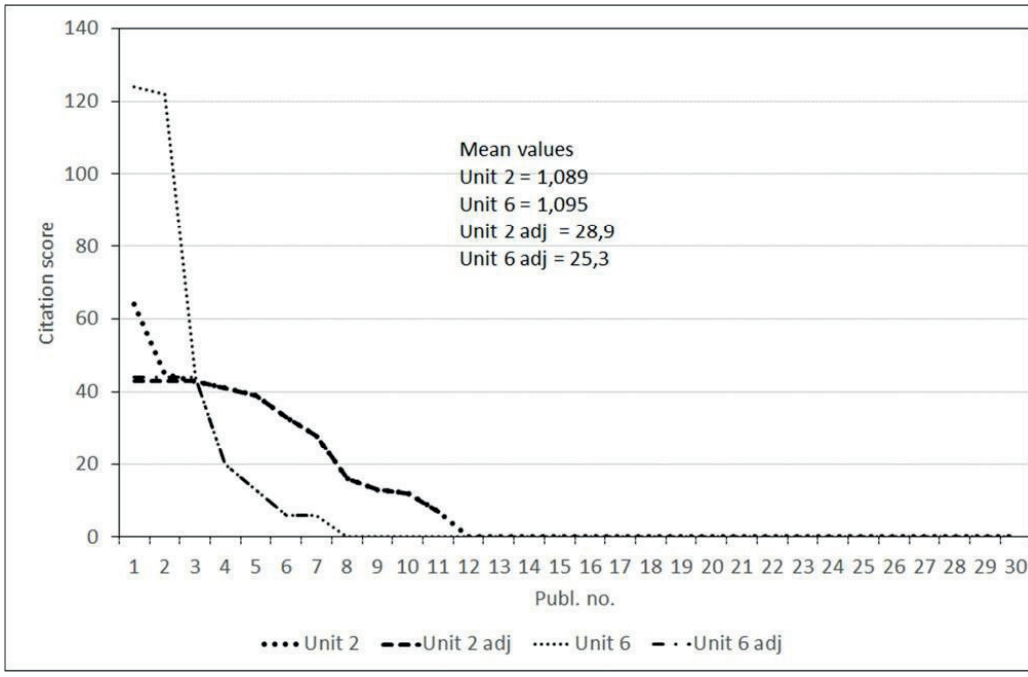


Figure 3. Distribution of citations of two selected research units in "Physics", source: Web of Science, update 2018.

A final question is why an adjusted mean is suggested instead of a Hirsch-Index for research units. The reason is that for the comparison of units in different fields of activity, a normalisation by the field averages is necessary. This implies that a normalisation by the Hirsch Index of the total field has to be calculated. However, the Hirsch Index is conceived for smaller samples of analysis. E.g., for the field of "Biotechnology & Applied
Microbiology", about 2900 publications appeared in 2015 (Figure 4). For this large sample, a Hirsch Index of 96 can be calculated which is far beyond the mean value of 9.4, thus it is unrealistic. In contrast, it is possible to determine an adjusted mean value. With 14.5, it is distinctly above the standard mean value, as about 50 percent of all publications have less than 6 citations. These publications are not included in the adjusted mean value. In this case, the cut

of the very high citations is less important for the level of the adjusted mean, although the first publication with high citations received 2335 citations in three years, the second 628, the third 559 and the fourth already 414. Thus, the first publication is not at all characteristic for the standard activities in this area and its exclusion in the adjusted calculation is justified.

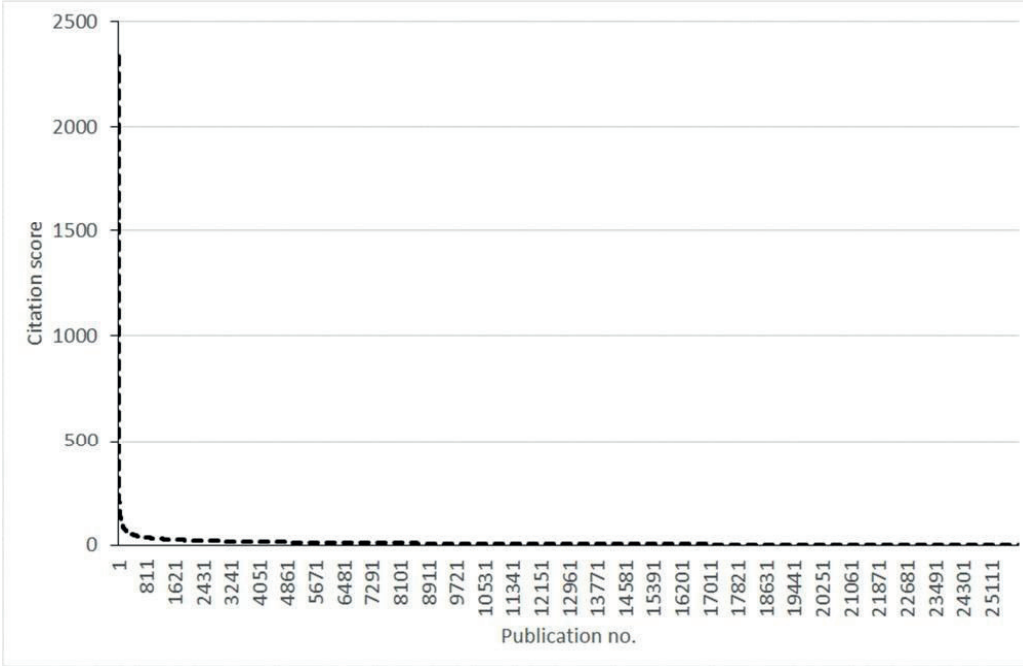


Figure 4. Distribution of citations of all publications of 2015 in the subject category "Biotechnology & Applied Microbiology", source: Web of Science, update 2018.

Conclusion

Nearly all distributions in bibliometrics are skew. In particular, the distribution of citations of publications by research units is skew. For a ranking of research units, it is recommended to replace the calculation of standard mean values by adjusted mean values where outliers with very high positive citations are cut and very those with low or no citations are cancelled. Such an adjusted mean value is oriented on the average activity of a research unit and leads to a more adequate assessment. This approach is based on the concept of the Hirsch-Index. This calculation often implies a different ranking of research units and it is important in cases where the distribution of finances depends on bibliometric rankings.

References

Adamic, L. A. (2002): Zipf, Power-laws, and Pareto - a ranking tutorial, retrieved January 2, 2019 from: http://www.labs.hp.com/research/idl/papers/ranking/ranking.html.

Chen, Y.-S. & Leimkuhler, F. F (1986): A relationship between Lotka's Law, Bradford's Law, and Zipf's Law, *Journal of the Association for Information Science and Technology*, 37(5), 307314.

De Solla Price (1976): A general theory of bibliometric and other cumulative advantage processes, *The Journal of the Assossiation for Information Science and Technology*, 27(5), 292-306.

Glänzel, W. & Moed, H. (2005): Journal impact measures in bibliometric research, *Scientometrics*, 53(2), 171–193 .

J. E. Hirsch, J. E. (2005): An index to quantify an individual's scientific research output, *PNAS*, 102 (46), 16569-16572.

Lotka, A. J. (1926): The frequency distribution of scientific productivity. *Journal of the Washington Academy of Sciences*, 16, 317–323.

Mandelbrot, B. (1959): A Note On a Class of Skew Distribution Functions: Analysis and Critique of a Paper by H. A. Simon, *Information and Control*, 2, 90-99.

Murphy, L. J. (1973): "Lotka's law in the humanities?", *Journal of the Association for Information Science and Technology*, 24(6), 461-462.

Narin, F. & Hamilton, K. S. (1996): Bibliomtric performance measures, Scientometrics, 36(3), 293-310.

Newman, M. E. J. (2005): Power laws, Pareto distributions and Zipf's law, *Contemporary Physics*, 46(5), 323-351.

Pao, M. L. (1986): An empirical examination of Lotka's Law, *Journal of the Association for Information Science and Technology*, 37(1), 26-33.

Pareto, V. (1935): The mind and Society, New York: Harcourt, Brace and Company.

Radhakrishnan, T. (1973): Lotka's Law and Computer Science Literature, *Journal of the Association for Information Science and Technology*, 24(6), 461-462.

Schmoch, U., Beckert, B. & Schaper-Rinkel, P. (2019): Impact assessment of a support programme of science-based emerging technologies, *Scientometrics*, forthcoming.

Seglen, O. (1992): The skewness of science, Journal of the Association for Information Science and Technology, 43(9): 628-638.

Simon, H. A. (1955): A Note On a Class of Skew Distribution Functions, *Biometrika,* 42(3/4), 5-440.

Zipf, G. H. (1949): *Human Behavior and the Principle of Least Effort*, Reading, Massachusetts: Addison-Wesley Press.

Using Internet Data to Complement Traditional Innovation Indicators

Lukas Pukelis[1] and Vilius Stanciauskas[2]

[1]*lukas.pukelis@ppmi.lt*
Public Policy and Management Institute (PPMI), Gedimino g. 50, 01110, Vilnius (Lithuania)

[2]*vilius@ppmi.lt*
Public Policy and Management Institute (PPMI), Gedimino g. 50, 01110, Vilnius (Lithuania)

Abstract
Currently, innovation activities of enterprises are assessed using either surveys or IPR/patent analysis. Neither of these methods are ideal for a variety of reasons, including high cost, high labour intensity, and certain problems with data reliability. This paper presents our efforts to develop new innovation indicators using internet data and supervised machine learning, to be used alongside the existing measures. The paper draws on the insights of the "Data4Impact" project, which sought to explore the possibilities of using Big Data methodologies to assess the impact of research funding. Initial results suggest that the method described in here, though not without its shortcomings, shows certain promise to capture innovation activities of enterprises. Its main advantages are the ability to capture non-patented innovations and to do so in a fraction of time compared to traditional survey approaches. This might benefit research funding bodies which seek to broaden their impact monitoring measures.

Introduction
Being able to assess and estimate the innovation activities of enterprises in the private sector is a highly salient question for scholars and policy-makers. Innovation is directly related to economic growth (Rosenberg 2004) and as such is one of the desired effects of the research-funding or certain economic policy measures. However, the tools available to scholars to measure the innovation among the private enterprises are limited and not ideal. Currently, it is mostly done by using innovation surveys or intellectual property (patent) analysis. Each of these methods when used in isolation or together can present a rich and detailed picture of innovation activities of private enterprises. However, they have several serious shortcomings that can impede their use. The main weakness of patent analysis is that patents do not directly correspond to innovations in a sense that: not all patented ideas become innovations; not all innovations are patented; propensity to patent differs among enterprises, and patents in different jurisdictions are not directly comparable (Archibugi & Planta 1996). Meanwhile, the main weakness of using surveys is the cost in terms of time and labour resources needed to carry them out. Currently, the most respected effort to carry out a wide-scale innovation survey in Europe is the "Community Innovation Survey" (CIS) by Eurostat. The main shortcoming of CIS data is that data collection, processing and publication can take up to four years (e.g. in 2018 most recent CIS data was from 2014). Even with smaller surveys the time lag between the initiation of the survey to the data can be considerable, often taking months and requiring hundreds of man-hours of labour. Particularly worrying recent trend is the declining survey response rates and survey fatigue, especially among the organisations benefitting from the EU research funding.

Given these shortcomings of existing methodologies for estimating innovations, there has been some recent interest of using internet data and other Big Data methodologies to derive additional indicators

to estimate innovations or other impacts of research funding. For instance, Centre for European Economic Research in Manheim University has explored using internet data to track the innovation activities of German enterprises (Kinne & Axenbeck 2018) and the European Commission has launched several initiatives to use Big Data methodologies to estimate the impact of research funding (EC 2015).

This paper presents partial results of one such project – "Data4Impact" funded under H2020 Co-Creation programme. While the project's overall goal was to use Big Data methods to provide a comprehensive summary of the outputs, results, and potential impacts of EU-funded projects, this paper concentrates on a single indicator – using internet data to estimate the number innovations produced by participating companies. The paper is structured as follows: the first part presents the overall rationale why company websites could be considered a valid data source and presents the pros and cons on using the internet data. The second part outlines the methodology for identifying innovation content in company websites and the key challenges of working with internet data. The third part briefly describes the results and benchmarks the company innovation counts from the internet to the 'hard' patent data.

Internet as a data source

There have been several attempts to use internet data to estimate innovations among companies in the private sector (Kinne & Axenbeck 2018). Nonetheless, the internet data is still considered novel and as such it poses certain validity and reliability concerns. Namely, there are two main questions regarding internet data or data from company websites: the *validity question* – do the indicators obtained from company websites correspond to the number of 'true' innovations; and *reliability* question – can we estimate the number of innovations in a reliable manner. The reliability question is purely technical in nature and is discussed in greater detail in the subsequent sections of the paper. Meanwhile, the validity question is more philosophical and can hardly be answered in a simple and straight-forward manner. Instead, we present an argument that the overall validity of the internet data is just as valid as the survey data.

First, over 85% of all enterprises in the EU have a website or some presence on the web[1]. Though the coverage is not universal, for all practical purposes it is large enough that we could consider that data on a certain enterprise could in principle be found and accessed on the web. Furthermore, there are no obvious biases in terms of countries, regions or economic sectors that could jeopardise the validity of the data.

Second, enterprises increasingly view web as an important platform to supply information about themselves and their activities. There are already some studies carried out that demonstrate that enterprises do post information on their innovation activities on the web and that overall it is broadly comparable to the innovation data from other sources (Gök, Waterworth & Shapira 2015; Katz & Cothey 2006; Youtie et al 2012; Aurora et al. 2013). Having said that, it is important to point out that there are certain peculiarities associated with using data from the company websites. It is important to note that the data on company websites is presented there to communicate the essential information about the enterprise in question to its target audiences, which might include: clients, business partners, and/or competitors. Scientists and policy analysts are not the target audience and the information on the websites is not tailored to their needs. This manifests in a variety of ways, for instance companies tend to focus on their products that are available for sale at the moment or will be in the immediate future. R&D activities with less immediate market applications tend to receive less coverage (Gök, Waterworth & Shapira 2015).

[1] Eurostat: Digital economy and society statistics – enterprises <https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Digital_economy_and_society_statistics_-_enterprises>

Furthermore, another significant issue arises from the fact that company website data is self-reported data. This means that different companies might have different propensity to announce their innovation activities or might apply different threshold to what exactly defines innovation (compared to e.g. incremental improvement). However, though definitions of what constitutes a genuine innovation might differ, enterprises do have a market incentive not to understate or overstate their innovation activities. Since companies with a propensity to under or over-sell their position would face negative reactions from partners and/consumers, enterprises are under certain pressure to provide accurate information on their webpages.

Though self-reporting this is a serious shortcoming, it is by no means unique to the internet data. Innovation surveys also rely on self-reporting and thus suffer from the same validity concerns. It is important to note that in surveys concerns over self-reporting are usually addressed by increasing the sample sizes. It is hoped that in a large sample different biases will cancel each other out and the indicators would be generally valid. However, in surveys ramping up the samples is often difficult and costly, whereas with internet data these marginal costs are practically non-existent and data collection from the web can be carried out for a fraction of the cost, while maintaining either larger samples or eliminating them altogether and carrying out full population studies.

Table 1 contains a summary of pros and cons of using different information sources to measure enterprise innovation. Compared to other sources, internet data has the advantage of being relatively cheap and rapid. This means that indicators computed using these data can be 'refreshed' more frequently and without major costs to either data collectors or enterprises themselves. However, that comes at a cost of lower granularity. As outlined in the next section, inferring innovation counts from company websites involves extensive natural language processing and working with free-text data. There is an enormous amount of variation in how companies structure their websites and content within. As such identifying innovation related content on company websites is a hard-enough task and classifying innovation content into smaller categories becomes increasingly complicated.

Table 1. Pros and cons of different sources to measure enterprise innovation

Method	*Pro*	*Con*
Survey	• Ability to directly ask desired questions directly to the target respondent; • Granularity – being able to go into detail	• Resource intensive; • Self-report bias; • Survey fatigue and declining response rates; • Recency bias – can only inquire about recent occurrences; • Low technological detail.
Patent/IPR analysis	• 'Hard data' – no self-reporting bias; • Ability to go back in time decades or centuries; • High technological detail.	• Fuzzy link between patents and innovations – not all innovations are patented and not all patents are innovations.
Internet data	• Cheap and rapid; • Possibility to have large samples/ full population studies; • Being able to go back in time months/years.	• Lower granularity; • Reusing data originally intended for different purpose and different audience.

Methodology

Defining and operationalising the concept of innovation for internet data

Oslo Manual considers innovation to be a continuous process, which can manifest through a variety of different types as product, process, marketing or organisational innovations (Oslo Manual 2018). Following the taxonomy of innovation activities of the Oslo Manual, the most reputable source of company innovation data in the EU – Community Innovation Survey also distinguishes between a variety of different innovation activity types. While, it is possible to collect data on the different types of innovation activities in the desired level of detail using the survey approach, it is not the case with internet data. Gathering data from the web, means that texts originally composed for different purposes are used to collect data on and measure company innovations. This in turn means that it is simply not possible to extract the same level of detail from these texts as is possible with the survey approach. As companies are entities with a clear goal to sell their products and services, they dedicate the biggest portion of their websites for describing them. Similarly, they can expose the user to the end product a marketing campaign without revealing any details on whether the campaign itself contained any marketing innovations.

Furthermore, as mentioned in the previous section, in their websites, companies tend to focus on innovation outputs – new products, services and improvements in the process rather then innovation activities with no immediate market application because these directly contribute to their core business. As such, in our methodology, we focussed solely on the innovation outputs and particularly on the "innovation announcement texts". We define "innovation announcement texts" as any massage in the company web domain that explicitly states that a company introduces a new product/service or improvement in the internal processes. By doing this, we are restricting our focus to only very explicit innovation announcements and risk not identifying such innovations that happened but were not explicitly presented as innovations. However, we consider this restriction necessary to ensure the providence and accuracy of our indicator. A few sample innovation announcement texts are presented in Table 2.

Table 2. Sample innovation announcement texts

CENTUM® VP R6.05 Integrated Production Control System - With a new processor module and an enhanced engineering function - CENTUM VP R6.05 Integrated Production Control System Yokogawa Electric Corporation (Tokyo: 6841) announces that it will release CENTUM® VP R6.05, an enhanced version its flagship integrated production control system, on October 24.	Systematic introduces new capabilities to SitaWare and IRIS solutions The new functions include 3D visualisation and will improve situational awareness, safety, and usability, among other benefits
Unigraf is introducing UCD-340, world's first integrated test equipment for testing DisplayPort™	Datalogic a g lobal leader in the automatic data capture and process automation markets, proudly announces release of IMPACT Software 12.0, the latest version of the well-known software by Datalogic for Vision Guided Robotics applications.

In other words, product (or service) innovations tend to leave an observable and detectable trace on company websites, whereas many other types of innovations do not. This is a core feature of using

internet data to measure innovations: the measurement is constrained to a segment of innovation activities – innovation outputs which are described and presented on the websites. In our measurement, we used a 2x2 matrix to differentiate between different kinds of innovation outputs. On one axis we distinguished between those products that are already on the market and can be bought or sold and those which are not on the market yet, such as products with scheduled launch date, drugs currently undergoing clinical trials or prototype/demonstrator versions of possible new products. On the other axis we differentiated between products that are tangible (have such physical characteristics as volume and mass) and those which are intangible – having no observable volume or mass. Products such as devices, tools, or consumables would be classified as tangibles, while software, databases or various services would fall into intangible category. This distinction is outlined in the innovation-output matrix below.

Data gathering and analysis

The methodology for gathering and analysing the data is presented in Figure 1. In essence, this method utilises supervised machine learning to recognise innovation related texts from company website and label them as such. For this method to work, it first requires a pre-labelled sample of innovation and non-innovation texts on which the machine learning model is trained. Prior to deploying the model, we have manually labelled texts from 500 enterprises and trained the model on the labelled sub-set. The model was trained to distinguish between two categories: innovation related text and non-innovation related items.

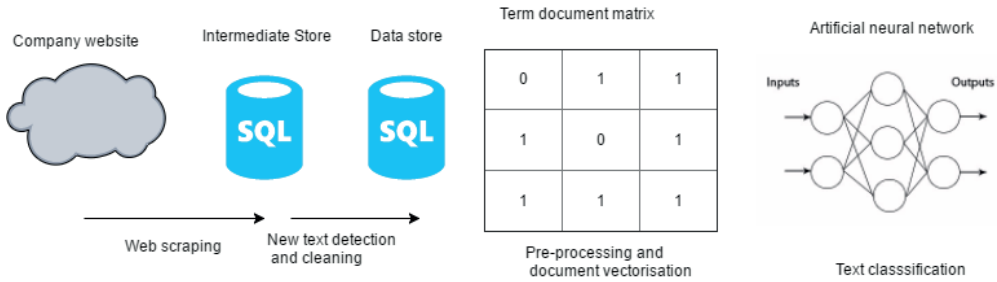


Figure 1. Workflow for harvesting and analysing internet data

Company innovation indicators are obtained in six steps:

1. Company websites are scraped and all textual data wherein is collected;
2. Data is placed in intermediate storage and new data is identified;
3. New data is indexed and placed in main data store;
4. Texts are pre-processed and vectorised;
5. Artificial network model is used to classify each text as innovation relate or not;
6. Company innovation scores are computed from aggregate data.

In the first step we crawled and scraped the company websites and collected text data wherein. To comply with the data protection regulation, we *ex ante* specified sections of websites which were ignored by the scraper, such as pages for contact information, biographies of company employees, etc. Where available instructions for scrapers located in *robots.txt* were obeyed. If an enterprise employed any kind of scraper blocks, they were obeyed and no measures were taken to counter them. Overall, we successfully scraped the websites of 1301 companies out of 1392 company sample, resulting in the response rate of 93,5%.

Scraped texts were placed in the intermediate data store. From there they were compared to the existing records (from previous scraping iterations) and new or partially new content was identified and placed for further analysis. Additionally, if some of the existing records were not found in the current iteration, they were marked as discontinued in the main database.
After the standard text pre-processing procedure (tokenization, lemmatization, stop-word and punctuation removal) texts were vectorized, i.e. converted to document term matrixes using '*doc2idx*' approach, which replaces text word tokens with corresponding numeric values, see Table 3.

Table 3. Illustration of doc2idx approach

Element	*Example*
Normal text	"A fool thinks himself to be wise, but a wise man knows himself to be a fool."
Tokens	'fool', 'thinks', 'be', 'wise', 'wise', 'man', 'knows', 'be', 'fool'
Dictionary	'fool': 1, 'thinks': 2, 'be': 3, 'wise': 4, 'man': 5, 'knows': 6
Vector	1,2,3,4,5,6,3,1

In this step we limited the size of dictionary to 50 000 entries. In developing the dictionary, we filtered out the extreme values: words that either appeared fewer than four times or words that appeared in more than 50% of the documents. After filtering the extremes, we selected 50 000 most common items. We arrived at 50 000 mark iteratively: our experiments with the labelled data suggested that the dictionary of this size is sufficient to get decent performance from the model, while ensuring that the model does not become unnecessarily large. We also standardized the texts to the fixed length of 1000 tokens. Texts shorter than the mark were padded with zero values, while longer texts were truncated, leaving first 1000 tokens. Similarly, 1000 token length was decided after reviewing the structural characteristics of the labelled sample, which revealed that 1000 token mark roughly corresponds to 95th percentile of the document length distribution (see Figure 2).

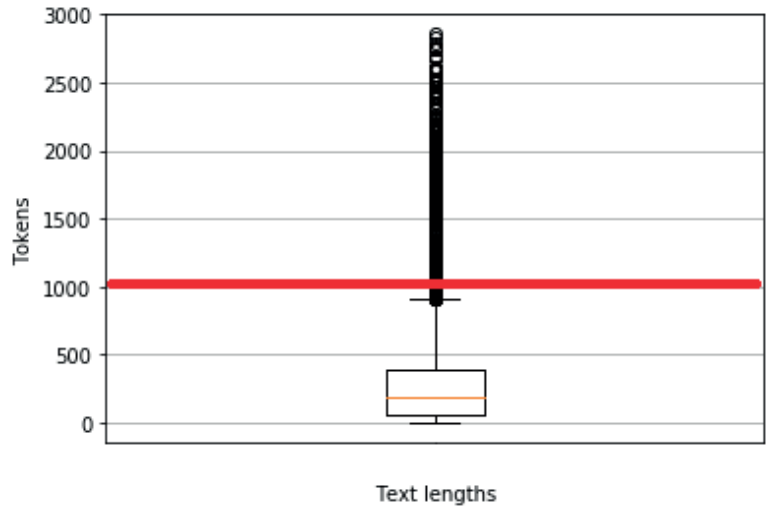


Figure 2. Distribution of texts by length

Once the document is vectorised, it can be put into the artificial neural network (ANN) model to be classified. These models take in vectorised documents and output a probability that a particular text is innovation-related. That way it is possible to take into consideration for how the word meanings might change on account of its context.
Overall, we chose the artificial neural network as the main model for text classification because these models can solve highly non-linear problems and, therefore, are almost uniquely suited to perform free-text classification tasks, where the complexity of data is extremely high. Prior to settling down on the text classification algorithm, we performed extensive testing, comparing the performance of the most commonly used classification algorithms (see Table 4).

Table 4. Comparison of different classification algorithms on the labelled data sample

N: 163 Companies; 31 898 webpages Training set: 23 923 (75%); Test set: 7 975 (25%)		*Logistic regression*		*Random forest*		*Support Vector Machine*		*ANN*	
		Model prediction		Model prediction		Model prediction		Model prediction	
		Neg.	Pos.	Neg.	Pos.	Neg.	Pos.	Neg.	Pos.
Actual Value	Neg.	7397	52	7403	46	7448	1	7442	7
	Pos.	166	360	159	367	448	78	41	485
F1 score:		0.767		0.781		0.257		0.952	

As indicated in Table 4, two other models: Logistic regression and Random Forrest also performed comparatively well, but their predictions were on average 20% less accurate than those of the ANN model. Most importantly, these models generated significant numbers of false negatives (model fails to identify innovation content), which is particularly problematic because, if need be, false positive predictions can be addressed and filtered out in later stages of the analysis, while false negatives mean that some innovation content remains permanently uncaptured.

We used a convolutional neural network (LeCun & Bengio 1995) because it evaluates not individual words, but phrases of specified length. More specifically, we constructed a convolutional neural network for text classification similar to the one describer by Kim (2014). The ANN model has a single dynamic embedding layer, followed by three parallel convolutional-dropout-pooling clusters with varying window sizes. These are followed by a concatation layer and two dense layers with a dropout layer in between. The overall schematic for the ANN model used is depicted in Figure 3.

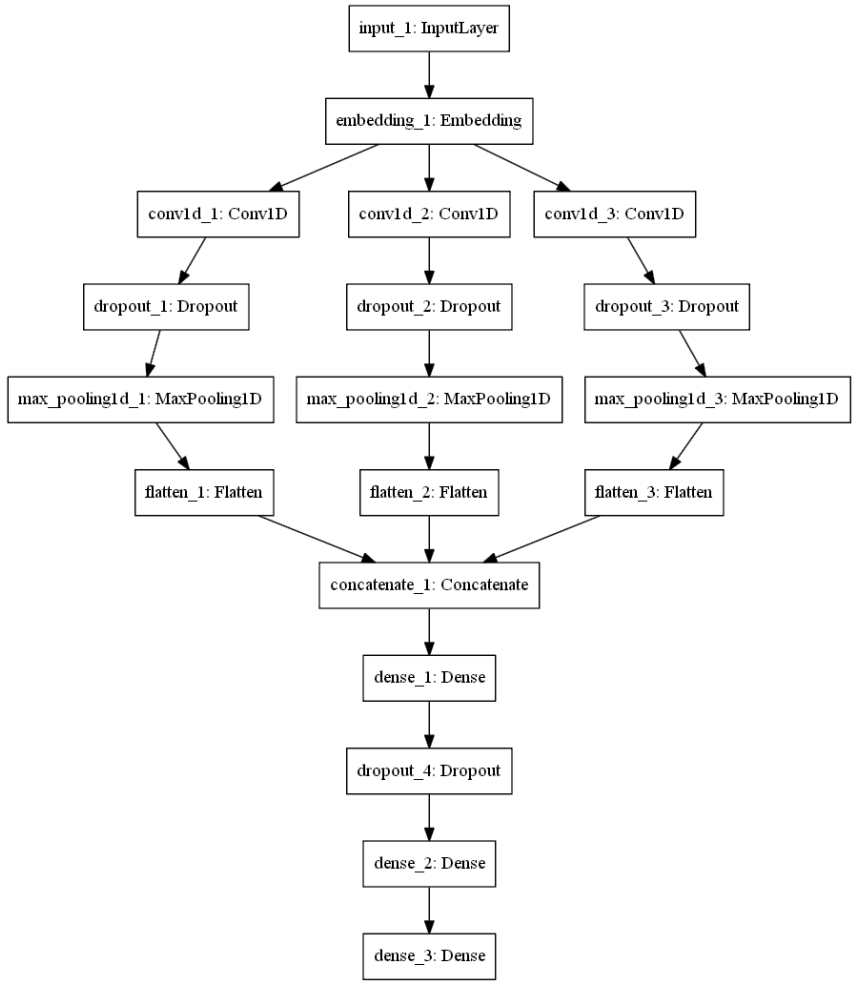


Figure 3. Schematic for the ANN model

After using the model to generate predictions for the whole dataset, we performed additional validation tests for model performance. A sample of 1 000 model predictions with equal balance between classes was drawn from the dataset these texts were manually checked and manual check assessment was compared with the model prediction. We discovered five cases of false positives and none false negatives. It was judged that the model performance is satisfactory and the model design was frozen with the same configuration in place.

Results

Share of Innovative companies

Our measurement yields that 588 companies out of 1301 participants have produced at least one innovation resulting in innovation rate of 45%, see Table 5. Though this figure is slightly smaller than the 'Share of Innovative Companies in EU-28' from CIS survey, the two figures are not directly comparable, as CIS utilises a much broader definition of innovation. If we look at the share of product innovative enterprises in CIS the figure is much smaller – just below 24% percent. Overall, out results

are consistent with the expectations that companies benefiting from EU research funding (as were the target group of the Data4Impact project) would have a higher share of innovating enterprises than the general enterprise population.

Table 5. Share of innovative companies and innovation content

	Records in DB	*Companies*	*CIS innovative companies*	*CIS product innovative companies*
All	567554	1301		
With Innovation	13815	588		
Share	2.5%	45%	49%	23.9%

Source: Own calculations; CIS 2014 data

Innovation mentions vs innovation counts

Table 5 shows the results at innovation mention and company level, which reveal the amount of innovation related web-pages in the sample and the count of the innovative companies in the population. However, what is missing is the number of innovations produced overall and by each innovative company. In other words, we need to translate innovation mentions into distinct innovation counts. This is far from trivial task, because while some companies dedicate multiple web-pages to a single product, other cram multiple products in a single web page.

To complete this task, we have developed a method which performs hierarchical clustering of innovation web-pages of the same enterprise. Cluster-groups where the distance between individual nodes is smaller than 0.5 are considered a single innovation, while cluster groups with a distance greater than 0.5 are considered distinct innovations. The cluster analysis is performed by selecting 10 most distinct (based on TFIDF scores) n-gram terms from each innovation web page and computing pairwise Jaccard distance between the company innovation pages. The 0.5 threshold was chosen through iterative manual validation.

Figure 4 shows the distribution of distinct innovations in our sample. As expected, majority of companies announced only a single innovation output, but is some cases it went as high as 20 innovation outputs. Overall, we identified distinct 785 innovation outputs for 588 innovative companies, averaging at around 1.4 distinct innovation output per innovative company.

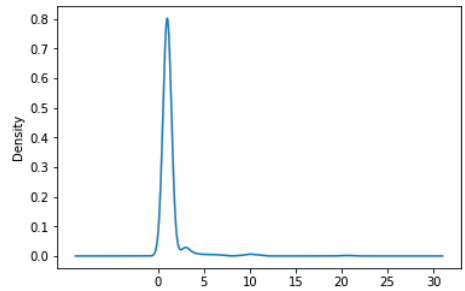


Figure 4. Kernel density plot for distinct company innovation counts

Innovation outputs by type

To assign innovation outputs to precise categories in the innovation output matrix (Table 2), we followed the same approach as with overall innovation output identification. We first developed a labelled set of innovation outputs assigning each innovation output to one or more specific categories. We then trained machine learning models to recognise different innovation outputs. The distribution of different innovation types is presented in Figure 5.

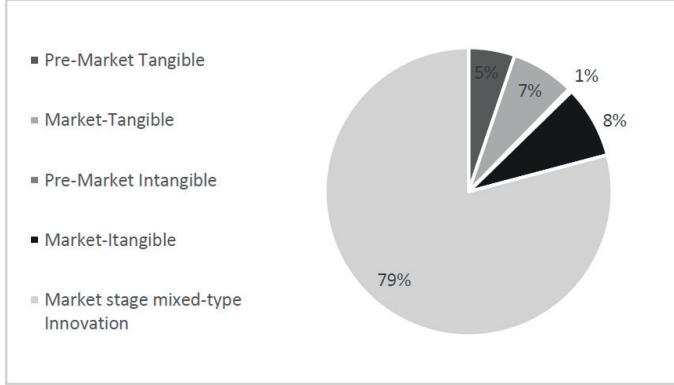


Figure 5. Innovation Outputs by Type

Most of the innovation outputs we identified were of mixed type, i.e. composite products falling into more than one category. This mostly corresponded to composite products with tangible and intangible components such as device AND software, drug AND therapy, infrastructure AND service.

However, it is important to note that our initial sample contained disbalanced classes with the share of certain innovation outputs far outpacing the other which has resulted in difficulty for the machine learning model to recognise certain innovation outputs categories.

Innovations from internet and patents

We compared our innovation output scores to other commonly used measures for company innovation, more specifically company patent counts. We compared the number of distinct company innovations to the overall number of patents and patent applications filed by the company. As shown in Figure 6 and Table 6, there is almost no correlation between the patent and web data.

Table 6. Correlation between Innovation Outputs and IPR

Pearson's r	0.0713

Source: Lens.org, own calculations

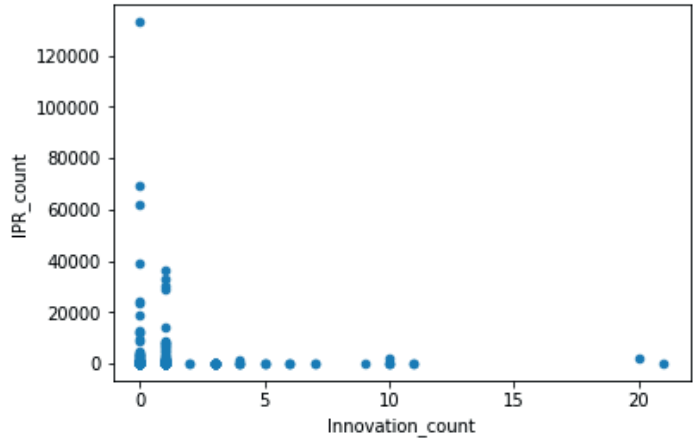


Figure 6. Innovation Outputs and IPR

Our findings closely echo previous studies on the topic, which also noted poor correspondence between patent and internet data (Gök, Waterworth, Shapira 2015). In addition to the aforementioned reasons why patents might not directly correspond to innovations (different motivations or propensity to patent, etc.), we must note two additional factors:

1. A significant number of innovations have a significant service component, which does not relate to patents in any way. For instance, an internet service provider might introduce an innovative service (e.g. 5G internet connection) without owning a single 5G-related patent.
2. In large multinational corporations a dedicated research division, which might have a significant number of patents might not have its own dedicated website (e.g. research arm of the Microsoft Corporation does not have its own website, only a subdomain in www.microsoft.com).

Overall, these findings show that internet data captures different aspects of innovation activities than the IPR indicators and, hence, can be used to complement them in future research.

Summary

This paper presented some of the results of the 'Data4Impact' project, aimed at utilising Big Data methodologies to estimate the impact of research funding. It summarised our experience in using internet as a source of company innovation data.

Overall, we consider that internet data, especially data from company websites, can be treated as a valuable source of company innovation data. Though further improvements to the data gathering and analysis methodology are needed, preliminary results indicate that mentions of innovation outputs can be captured from the company websites with high degree of accuracy using supervised machine learning techniques. However, as this is one of the first attempts to leverage internet data to capture company innovation, it naturally needs more fine-tuning and refinement before it can reach its full potential.

Funding acknowledgements:

This work is based on research funded by the European Union's Horizon 2020 research and innovation programme under grant agreement No 770531.

References:

Archibugi, D. and Planta, M., 1996. Measuring technological change through patents and innovation surveys. Technovation, 16(9), pp.451-519.

Arora, S.K., Youtie, J., Shapira, P., Gao, L. and Ma, T., 2013. Entry strategies in an emerging technology: a pilot web-based study of graphene firms. *Scientometrics*, *95*(3), pp.1189-1207.

European Commission, 2015. Big Data Can Show Research Impact. Horizon: EU research and innovation magazine. < https://horizon-magazine.eu/article/big-data-can-help-show-impact-research-projects-eu-commissioner-moedas.html>

Gök, A., Waterworth, A. and Shapira, P., 2015. Use of web mining in studying innovation. Scientometrics, 102(1), pp.653-671.

Katz, J.S. and Cothey, V., 2006. Web indicators for complex innovation systems. Research Evaluation, 15(2), pp.85-95.

Kim, Y., 2014. Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882*.

Kinne, J. and Axenbeck, J., 2018. Web mining of firm websites: A framework for web scraping and a pilot study for Germany. ZEW-Centre for European Economic Research Discussion Paper, (18-033).

LeCun, Y. and Bengio, Y., 1995. Convolutional networks for images, speech, and time series. *The handbook of brain theory and neural networks*, *3361*(10), p.1995.

Oslo Manual., 2018. Proposed guidelines for collecting and interpreting technological innovation data. OCDE: Statistical Office of the European Communities.

Rosenberg, N., 2004. Innovation and economic growth. Innovation and Economic Growth, 52.

Youtie, J., Hicks, D., Shapira, P. and Horsley, T., 2012. Pathways from discovery to commercialisation: using web sources to track small and medium-sized enterprise strategies in emerging nanotechnologies. *Technology Analysis & Strategic Management*, *24*(10), pp.981-995.

What share of researchers publish monographs?

Emanuel Kulczycki[1] and Przemysław Korytkowski[2]

[1] *emek@amu.edu.pl*
Scholarly Communication Research Group, Adam Mickiewicz University in Poznań (Poland)

[2] *pkorytkowski@zut.edu.pl*
West Pomeranian University of Technology in Szczecin (Poland)

Abstract

In this study, we investigate what share of researchers publish monographs across fields, gender and seniority. We acquire data from the Polish current research information system, containing metadata about all publications by 67,415 Polish researchers, including 30,185 monographs and 638,779 articles from 2013-2016. The data are aggregated at the researcher level which allow us to shed new light on publication patterns in all fields, especially on monograph patterns which in previous studies have been investigated mostly in only the social sciences and humanities. The key finding of our study is twofold. Firstly, we show that scholars who publish monographs also publish journal articles at the same time. This pattern is observed in all dimensions, i.e. fields, gender and seniority. However, substantial differences between the fields are observed. Secondly, presenting the publication patterns at the researcher level allows us to argue that a monograph is the key publication channel for social sciences and humanities. The discussion summarizes our empirical findings and positions them in the light of other methods of data aggregation.

Introduction

Monographs are the primary type of scholarly book publication in social sciences and humanities (Giménez-Toledo & Román-Román, 2009; Williams, Stevenson, Nicholas, Watkinson, & Rowlands, 2009). The so-called hard sciences also use monographs, but it is definitely a less popular form of scholarly communication channel (Aagaard, Bloch, & Schneider, 2015; Kulczycki, 2018).

Systematic empirical studies regarding the share of monographs among scholarly publications are rare because full publication records are often not available. The most comprehensive studies have to be based on national level databases (Sīle et al., 2018). Those studies are conducted at the level of countries, fields or disciplines. Engels et al. (2018) show that scholarly book publications are not disappearing from scholarly communications, and the publication patterns seem stable over the last several years. Kulczycki et al. (2018) present publication patterns in eight European countries and argue that monograph patterns differ both across fields and across countries within social sciences and humanities (SSH). On average, the share of monographs in the total number of publications in those countries was 6.2% in 2014. The highest share of monographs in the total volume was observed in the Czech Republic (12.8%) with the lowest in Flanders (Belgium) at 1.8%. Also, at the level of disciplines, patterns were different. For instance, in the field of Law, the highest share was in Finland (11.1%) and the lowest in Flanders (2.28%). Such studies at field and country levels provide important information on how monographs are popular within fields and given countries, yet they do not detail how single researchers use monographs as a publication channel nor how often they decide to use it to communicate their research results. Moreover, existing literature devoted to monographs is most often limited to SSH fields.

Claiming that monographs are the primary research output in some fields, especially in SSH, has not been extensively documented or investigated. Moreover, analyses are usually conducted at the level of fields. Studies at the researcher level are most often related to citation analysis (Aksnes, Rorstad, Piro, & Sivertsen, 2011) and evaluation context at the national level (Sivertsen, 2016a) rather than the publication channel. Therefore, there is a lack of studies at the researcher level related to publication patterns. In this light, Sivertsen's study on patterns of

internationalization in the SSH performed at the researcher level reveals that 15.7% of researchers in humanities and 8.5% researchers of researchers from social sciences published a monograph within the studied four-year period (Sivertsen, 2016b).

In order to perform publication pattern analysis at the researcher level in a given country, a current research information system with metadata about all publications by all researchers is required. Therefore, such studies cannot be based either on the Web of Science or Scopus as these databases do not include all publication channels, especially local journals, monographs, and edited volumes. Current research information systems at the national level exist in Norway, Finland, Slovenia, and Poland, among others.

In this paper we fill this gap in research literature related to monograph patterns in all fields of science by investigating empirically what share of researchers publish monographs in a four-year period using comprehensive publication data collected in Poland. We also analyze what share of researchers in a given field published at least one or more monographs across both gender and seniority. Moreover, we analyze how publishing monographs is related to publishing journal articles. The discussion summarizes our empirical findings and positions them in the light of the often raised statement that monographs are the primary research output in SSH.

Data and methods

In this study we use bibliographical records of publications from the Polish Scholarly Bibliography, a part of the Polish current research information system POL-on, published for the years 2013–2016. Since 2013, all Polish higher education institutions and research institutes have been obliged to submit bibliographical records of publications affiliated with those institutions. We have data on 67,415 Polish researchers and their 30,185 monographs and 638,779 articles. The POL-on data contains publications by 87,352 researchers employed by Polish universities, basic and applied sciences institutes. However, in our analysis, we limited this number to include only those researchers who; (1) were academic staff member in the higher education institutions or research institutes; (2) published at least one publication of any type in the 2013–2016 period according to the POL-on data, and (3) obtained a PhD degree before 2013. In the final data set of 67,415 researchers, 273 (0.4%) and 1,584 (2.4%) of these researchers did not assign information on gender or the years of their PhD, respectively.

The POL-on data is originally aggregated at the researcher level, meaning that whole counting is used: every co-author gets credit for a whole publication.

We assigned information on gender using first name dictionaries. Moreover, we used information about the date of obtaining their PhD and the discipline attributed by the researcher. For 500 researchers, there were data about more than one discipline. We decided to exclude those researchers from the analysis and they were not counted in the number of 67,415 researchers. We mapped these disciplines to the six major fields of science and technology classification of the Organisation for Economic Co-operation and Development (OECD, 2007). In this way all publications published by researchers classified for instance as philosophers, are counted in this analysis as publications from Humanities.

Results

In each OECD major field, the majority of scholars did not publish any monograph in the studied four-year period. In Humanities, the share of such non-publishing scholars is 50.5% (4,347 researchers) and in Medical and Health sciences 92% (11,424 scholars).

The univariate logistic regression analysis of the studied factors is presented in Table 1. Researchers from Natural sciences were taken as the reference point in the logistic regression. Only researchers from Medical and Health sciences write fewer monographs than from Natural sciences (OR = 0.665). It is no surprise that researchers from SSH write many more monographs (OR = 6.424 and OR = 7.549, respectively). Male researchers (57% of all researchers), then

researchers from the hard sciences, by a small margin, write more monographs than female researchers (OR = 1.146).

Table 1. Univariate logistic regression of researchers who published and did not publish monographs across the major OECD fields, gender and seniority.

Parameter	N	Researchers who did not publish a monograph	Researchers who published monograph(s)	Odds ratio	95% confidence interval	p-value
OECD Major Field						
Natural sciences	13,786	12,201 (88.5 %)	1,585 (11.5 %)	1	(Ref.)	
Engineering and technology	13,769	10,697 (77.7 %)	3,072 (22.3 %)	2.211	(2.07; 2.362)	< 0.001
Medical and Health sciences	12,411	11,424 (92 %)	987 (8 %)	0.665	(0.612; 0.723)	< 0.001
Agricultural sciences	4,418	3,559 (80.6 %)	859 (19.4 %)	1.858	(1.696; 2.035)	< 0.001
Social sciences	14,421	7,861 (54.5 %)	6,560 (45.5 %)	6.424	(6.041; 6.835)	< 0.001
Humanities	8,610	4,347 (50.5 %)	4,263 (49.5 %)	7.549	(7.059; 8.076)	< 0.001
Gender						
Female	28,809	21,824 (75.8 %)	6,985 (24.2 %)	1	(Ref.)	
Male	38,333	28,046 (73.2 %)	10,287 (26.8 %)	1.146	(1.106; 1.187)	< 0.001
Seniority (number of years after PhD)						
0–10	35,334	26,874 (76.1 %)	8,460 (23.9 %)	1	(Ref.)	
11–20	15,966	11,236 (70.4 %)	4,730 (29.6 %)	1.337	(1.282; 1.394)	< 0.001
21–30	7,247	5,315 (73.3 %)	1,932 (26.7 %)	1.155	(1.09; 1.223)	< 0.001
30+	7,284	5,520 (75.8 %)	1,764 (24.2 %)	1.015	(0.957; 1.077)	0.617

Figure 1 shows how many monographs researchers published across OECD major fields, and how publishing monographs is related to publishing journal articles. In all fields, the overwhelming majority of researchers who published monographs also published articles. 74.30% of researchers (all fields) did not publish a monograph, but even then, the majority of them published at least 3 articles in the four years period.

Analysing the all fields together, 16.52% of researchers published one monograph, 5.30% published two monographs, and 3.89% published three or more monographs. These numbers differ when the analysis is broken down into the major fields. For instance, 8.34% of researchers in Humanities published three or more monographs, whereas just 0.77% of researchers in Medical and Health Sciences did so.

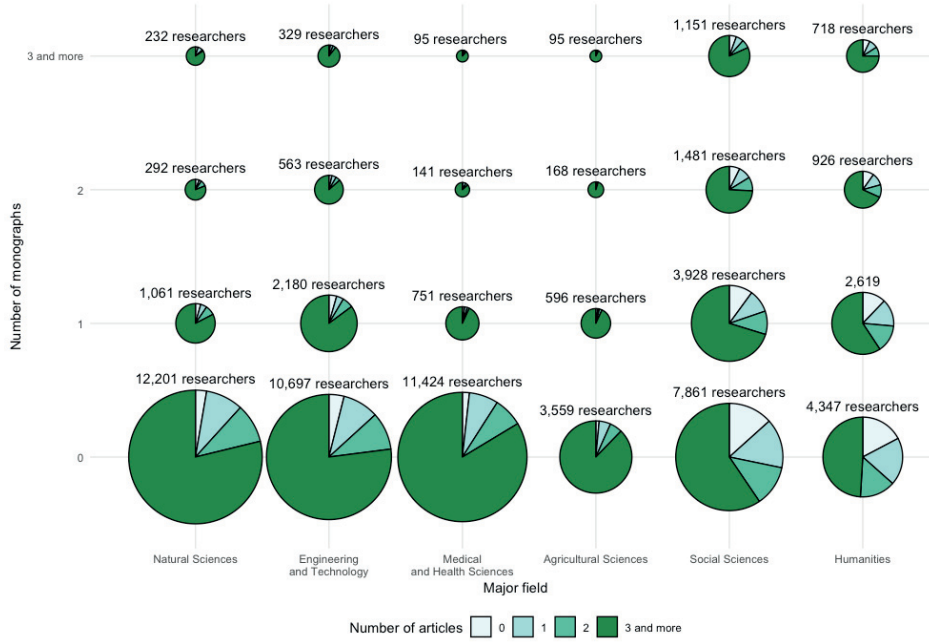


Figure 1. Publication patterns across the major OECD fields, with the number of published monographs and articles.

Figure 2 shows the differences between the number of published monographs across the major OECD fields and gender. In all fields except Humanities, the majority of female and male scholars did not publish a monograph.

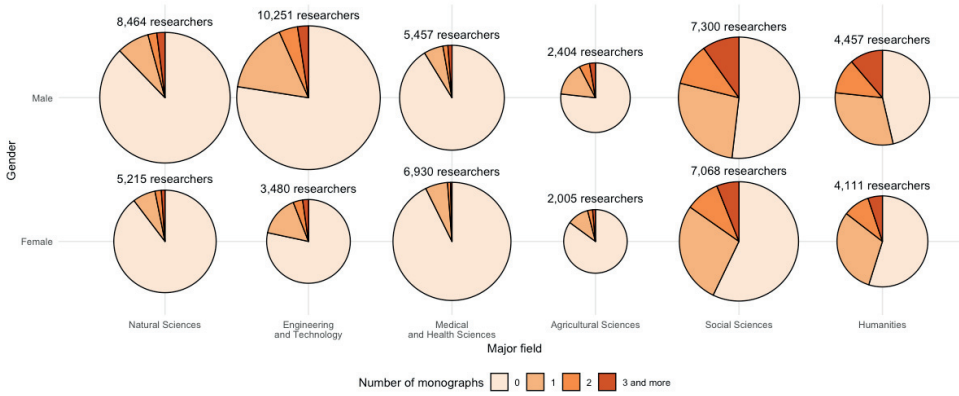


Figure 2. Publication patterns across the major OECD fields, with the number of published monographs and gender.

Discussion and conclusion

In this study we analysed the share of researchers who published monographs in the four-year period 2013-2016 across the major OECD fields, gender and seniority. We studied publications by over 67 thousand Polish researchers. The key finding of our study is twofold. Firstly,

scholars who published monographs also published journal articles at the same time. This pattern is observed in all dimensions, i.e. fields, gender, and seniority. Secondly, presenting the publication patterns at the researcher level allows us to provide stronger arguments that a monograph is the primary publication channel for social sciences and humanities.

Our study reveals that there is no substantial difference between female and male researchers in terms of writing at least one monograph in the four-year period. Virtually all those researchers published more scientific articles than monographs, with the majority of those researchers publishing both articles and monographs. We believe that this should be interpreted in such a way that working on a monograph does not mean abandoning scholarly communication using journal articles. The seniority of the researcher (measured as the number of years after obtaining their PhD) did not significantly change the inclination toward publishing monographs.

The number of monographs published by Polish researchers might be influenced by a recent practice in academic promotion procedures where publishing monographs was perceived as the proper way of obtaining the habilitation degree and professorship in many fields. For instance, in 2011–2016, less than 2% of researchers from chemical and pharmaceutical sciences published a monograph as their habilitation thesis, whereas in all SSH fields this share was above 70% (Kulczycki, 2019).

Comparing the Polish data on SSH (from 2013–2016) with the Norwegian data (2010–2013) presented by Sivertsen (2016b), substantial differences can be observed between the share of researchers who published a monograph. According to Sivertsen, 8.5% of Norwegian researchers from social sciences published a monograph compared to 45.5% of Polish researchers. A similar situation can be observed in humanities, with 15.7% of Norwegian and 49.5% of Polish researchers publishing a monograph. As Kulczycki et al. (2018) show, publication patterns differ both across fields and across countries, e.g. publication patterns for humanities in Norway differ from those in Poland. Our results are line with this conclusion. However, we would like to emphasize the differences between publication patterns when the data are aggregated at the researcher level instead of the field level.

The monograph is perceived as the primary research output, especially in social sciences and humanities (SSH). When we look at the four-year period 2013-2016, the majority of Polish researchers did not publish a monograph, while for humanities this share is 50.5%. On one hand, it could be said that a four-year period is too short to analyze such a publication pattern, and on the other hand, we can say that in this specific Polish case, the four-year period (2013–2016) comprises the entire evaluation period counted for Polish higher education institutions and research institutions, motivating all researchers to present their best publications within this regular period to maintain an adequate evaluation.

Most previous studies on publication patterns focused on differences between fields, so publications were aggregated at the level of fields. In our study, we aggregated the publications at the level of researcher. For instance, Kulczycki et al. (2018) shows that monographs constituted 9.98% of all publication types in SSH published by Polish researchers in the four-year period 2011–2014. However, our study shows that in the four-year period 2013–2016, 47.0% of SSH researchers (two fields counted as a whole) published a monograph. A similar situation is in Norway where, according to Kulczycki et al. (2018), monographs constitute 3.56% of all publication types in SSH, whereas according to Sivertsen (2016b), 11.1% of all SSH researchers (two fields counted as a whole) published a monograph. This means that researchers use many publication channels at the same time to communicate research results. Counting publications at only the field level does not allow a proper emphasis on the role of monographs in the so-called soft sciences. Thus, we argue that presenting the publication patterns at the researcher level provides a stronger argument that the monograph is the primary publication channel for SSH.

Acknowledgments

This work was supported by the DIALOG Program (grant name "Research into Excellence Patterns in Science and Art"). The authors would like to thank Ewa A. Rozkosz for her support.

References

Aagaard, K., Bloch, C., & Schneider, J. W. (2015). Impacts of performance-based research funding systems: The case of the Norwegian Publication Indicator. *Research Evaluation*, *24*(2), 106–117. http://doi.org/10.1093/reseval/rvv003

Aksnes, D. W., Rorstad, K., Piro, F., & Sivertsen, G. (2011). Are female researchers less cited? A large-scale study of Norwegian scientists. *Journal of the American Society for Information Science and Technology*, *62*(4), 628–636. http://doi.org/10.1002/asi.21486

Engels, T. C. E., Istenič Starčič, A., Kulczycki, E., Pölönen, J., & Sivertsen, G. (2018). Are book publications disappearing from scholarly communication in the social sciences and humanities? *Aslib Journal of Information Management*, AJIM-05-2018-0127. http://doi.org/10.1108/AJIM-05-2018-0127

Giménez-Toledo, E., & Román-Román, A. (2009). Assessment of humanities and social sciences monographs through their publishers: A review and a study towards a model of evaluation. *Research Evaluation*, *18*(3), 201–213. http://doi.org/10.3152/095820209X471986

Kulczycki, E. (2018). The diversity of monographs: Changing landscape of book evaluation in Poland. *Aslib Journal of Information Management*, *70*(6), 608–622. http://doi.org/10.1108/AJIM-03-2018-0062

Kulczycki, E. (2019). Field patterns of scientometric indicators use for presenting research portfolio for assessment. *Research Evaluation*. http://doi.org/10.1093/reseval/rvy043

Kulczycki, E., Engels, T. C. E., Pölönen, J., Bruun, K., Dušková, M., Guns, R., … Zuccala, A. (2018). Publication patterns in the social sciences and humanities: evidence from eight European countries. *Scientometrics*. http://doi.org/10.1007/s11192-018-2711-0

Organisation for Economic Co-Operation and Development (OECD). (2007). *Revised field of science and technology (FOS) classification in the frascati manual DSTI/EAS/STP/NESTI(2006)19/FINAL*. Retrieved from http://www.oecd.org/science/inno/38235147.pdf

Sīle, L., Pölönen, J., Sivertsen, G., Guns, R., Engels, T. C. E. E., Arefiev, P., … Teitelbaum, R. (2018). Comprehensiveness of national bibliographic databases for social sciences and humanities: Findings from a European survey. *Research Evaluation*, (June), 1–13. http://doi.org/10.1093/reseval/rvy016

Sivertsen, G. (2016a). A bibliometric indicator with a balanced representation of all fields. In *International Conference on Science and Technology Indicators (STI 2016)* (pp. 910–914). Valencia, Spain.

Sivertsen, G. (2016b). Patterns of internationalization and criteria for research assessment in the social sciences and humanities. *Scientometrics*, *107*(2), 357–368. http://doi.org/10.1007/s11192-016-1845-1

Williams, P., Stevenson, I., Nicholas, D., Watkinson, A., & Rowlands, I. (2009). The role and future of the monograph in arts and humanities research. *Aslib Proceedings*, *61*(1), 67–82. http://doi.org/10.1108/00012530910932294

Cui Prodest? Reciprocity of collaboration measured by Russian Index of Science Citation

Vladimir Pislyakov[1], Olga Moskaleva[2] and Mark Akoev[3]

[1] *pislyakov@hse.ru*
Library, National Research University Higher School of Economics, Myasnitskaya 20, Moscow, 101000 (Russia)

[2] *o.moskaleva@spbu.ru*
Saint Petersburg State University, Universitetskaya emb. 7-9, St. Petersburg, 199034 (Russia)

[3] *m.a.akoev@urfu.ru*
Ural Federal University, Mira str. 19, Ekaterinburg, 620002 (Russia)

Abstract

Scientific collaboration is often not perfectly reciprocal. Scientifically strong countries/institutions/laboratories may help their less prominent partners with leading scholars, or finance, or other resources. What is interesting in such type of collaboration is that (1) it may be measured by bibliometrics and (2) it may shed more light on the scholarly level of both collaborating organizations themselves. In this sense measuring institutions in collaboration sometimes may tell more than attempts to assess them as stand-alone organizations. Evaluation of collaborative patterns was explained in detail, for example, by Glänzel (2001; 2003). Here we combine these methods with a new one, made available by separating 'the best' journals from 'others' on the same platform of Russian Index of Science Citation (RISC). Such sub-universes of journals from 'different leagues' provide additional methods to study how collaboration influences the quality of papers published by organizations.

Introduction

This paper is a next stage in deep analysis of a relatively new national citation database RISC (Russian Index of Science Citation, launched in 2005). In the article by Moskaleva et al. (2018) the main characteristics and internal structure of RISC were for the first time presented to international audience. Next, our conference paper (Akoev, Moskaleva & Pislyakov, 2018) discussed the essential findings about publication profiles of Russian organizations obtained from the RISC database. Rather intriguing facts were revealed, but it was an analysis of universities/research institutes as stand-alone units.

Here we explore collaboration between higher education institutions, their interdependence and their roles in this interdependence.

We briefly summarize here for a reader the results of our previous works which are essential for this paper and present new findings on collaborative performance of universities.

What is essential and motivated us in this series of studies:

(1) the 'segregation by databases' method sometimes is more tangible than citation analysis. The latter, of course, has theoretically greater potential but it needs longer time lag and more citation database accuracy to be reliable;

(2) these methods are not focused on a single nation case (Russia) but may be used for world literature as soon as 'publication sub-universes' are already established by such sub-databases as other national indexes on the Web of Science (WoS) platform or even WoS Emerging Sources Citation Index as opposed to SCIE+SSCI+AHCI, 'main journal indexes'. One example of this international technique will be presented later in our paper.

RISC and RISC Core

Russian Index of Science Citation (RISC) was launched in 2005 on the platform eLIBRARY.RU of the Russian company *Scientific Electronic Library*. As of January 2019, RISC contains more than 31 mln documents with more than 360 mln references, the total list of journals contains more than 60,000 titles, about 6,000 of them are fully indexed in RISC, other journals are sources of documents with Russian affiliations and papers citing them. Along with journals, RISC indexes conference proceedings, books, patents, dissertations and other research artefacts.

As it follows from its name, RISC main feature is citation indexing and analytics. However, in our study of collaborations we will not use citation analysis. What is more important for the present paper is that in 2017 a special subset of journals was defined, RISC Core. It contains those RISC journals which are also indexed either in Web of Science or in Scopus. It is a kind of 'best journals' subset of RISC periodicals.

It should be mentioned that since 2016 there is a special national citation index on the Web of Science platform, Russian Science Citation Index with about 800 journals (RSCI, not to be confused with RISC which is a much bigger database on the separate eLIBRARY.RU platform, with no connection to WoS). Journals for RSCI were selected through professional expertise by specially summoned Russian academia board and by expert opinion of top Russian scientists, more than 12,500 of them voted for journals. All RSCI journals are also included into RISC Core subset. It is important as soon as makes RISC Core independent from decisions made by foreign commercial companies Clarivate Analytics and Elsevier on inclusion/non-inclusion of Russian titles into their databases. This adds value to RISC Core as an instrument for analysis of Russian organizations' performance.

For assessment of the RISC Core a brief experiment was made. It was found that RISC Core which contained in 2016 only 23% of papers of all eLIBRARY.RU platform has attracted 93% of all citations there. This affirms the quality of the RISC Core sources and their status of, so to say, 'best journals'.

Further details on RISC may be found in (Moskaleva et al., 2018; Akoev, Moskaleva & Pislyakov, 2018).

Data and overview of previous results

We study 16 Russian universities and their output indicators taken from RISC database. These universities were chosen by three criteria:

(1) they should be leaders by number of documents in RISC as a whole;
(2) they should be leaders by number of documents in RISC Core;
(3) they should be involved into collaboration with the most of other chosen universities.

In tables and graphs we use abbreviated names of these institutions, but in the text we will usually spell them out. The full list of the universities with their abbreviations may be found in Appendix 1. We omit those collaborations which resulted in less than 20 papers in the whole RISC and regard these data as insufficient in all tables.

Output indicators of these organizations for the whole RISC and for the RISC Core are taken for five-year 2013–2017 period from the predefined reports in their profiles at eLIBRARY.RU.

The same approach was used by Akoev, Moskaleva and Pislyakov (2018) for 2012–2016 time period. Here we reproduce a graph from there (Figure 4 there, Figure 1 here), most important for the current analysis. The graph is corrected so that only universities studied in this paper are shown.

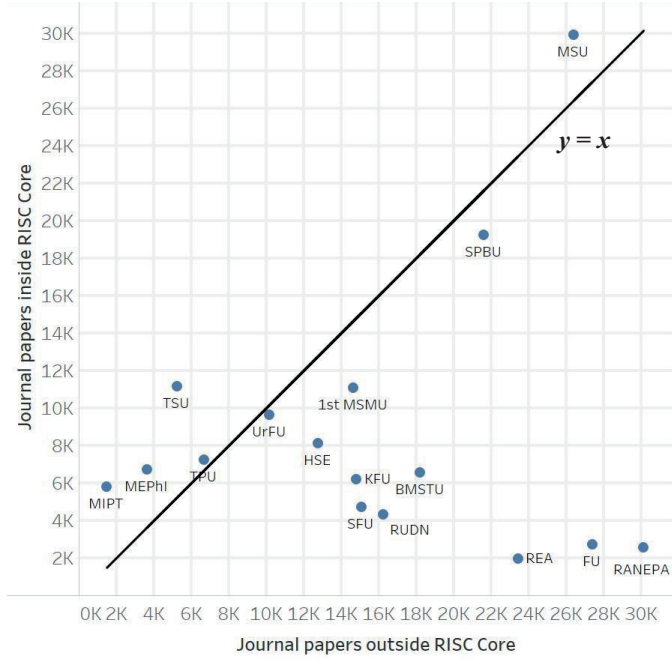


Figure 1. Most prolific Russian Universities: journal papers inside and outside RISC Core. (adapted from Akoev, Moskaleva and Pislyakov, 2018)

We can see that for different universities the share of documents in the Core is drastically different. The highest percentage of the Core papers is found for Moscow Institute of Physics and Technology (MIPT), Moscow Engineering Physics Institute (MEPhI) and Tomsk State University (TSU). The lowest share of their papers in 'the best journals' have Academy of National Economy (RANEPA), Plekhanov University of Economics (REA) and Financial University (FU). This previous analysis of the universities as stand-alone organizations will be useful in our study on collaborations.

Results and discussion

Symmetrical (bi-directional) collaboration indexes

First, we study simple symmetrical indexes measuring the strength of collaboration between Russian universities. Generally, two such indexes may be used, Jaccard index or Salton measure (Glänzel, 2003). We use here Jaccard and calculate it both for collaboration in the whole RISC and only in RISC Core.

Jaccard index is a ratio of the cardinality of the intersection of two sets to the cardinality of the union of these sets, expressed as a percentage. In our case, it is the number of papers

written by both organizations jointly divided by the number of papers written at least by one of these organizations. This index gives an estimate of the relative strength of collaboration between two organizations, normalized in a way by the volume of their outputs. Five joint papers are a weaker level of the collaboration when each of the institutions has 1000 articles than when they have published only 10 papers each. Jaccard aims to correct for this difference.

To illustrate, the highest Jaccard index in our data is found for collaboration of two Tomsk universities (State and Polytechnic, TSU and TPU) in the RISC Core — 7.22%. This means that every 14th paper (100/7.22) written by scientists *either* from TSU *or* TPU in the RISC Core journals is *jointly* written by them. The full data on Jaccard indices may be found in Appendix 2.

For our study, the Jaccard indices themselves are not so interesting as their ratio for collaboration in the RISC Core and RISC as a whole. It will reveal the change of the importance of collaboration between two organizations when moving from the entire database to 'the best journals'. Table 1 shows Jaccards for inside-core collaboration of each pair of universities divided by Jaccard for their co-authorship in the whole RISC. As soon as Jaccard is a symmetrical indicator, the matrix in Table 1 is also symmetrical.

Table 1. Ratio of Jaccard indices for collaboration in RISC Core/whole RISC. (light blue cell: <1; bold font: >2; yellow cell: >3; empty cell: insufficient data)

	MSU	FU	SPBU	RANEPA	REA	UrFU	SFU	RUDN	BMSTU	KFU	TSU	1st MSMU	HSE	TPU	MEPhI	MIPT
MSU		0.81	1.80	1.13	0.82	**2.06**	1.54	1.12	1.87	1.97	**2.14**	1.61	1.49	**2.09**	**2.33**	**2.16**
FU	0.81		1.24	1.04	1.97	0.66	1.62	1.08	1.94	**2.66**		**4.65**	1.91		**2.43**	**2.71**
SPBU	1.80	1.24		1.02	1.01	1.85	1.77	1.32	**2.41**	**2.32**	**2.08**	1.28	1.43	1.26	**2.47**	**2.42**
RANEPA	1.13	1.04	1.02		1.31	0.55	0.64	**2.27**	0.98	0.82	0.23	1.38	**2.55**		1.24	
REA	0.82	1.97	1.01	1.31		0.68	1.26	**2.82**	1.67	**3.23**		1.28	0.99		**3.38**	
UrFU	**2.06**	0.66	1.85	0.55	0.68		**2.34**			**2.26**	1.46		1.26	**2.08**	1.79	**2.10**
SFU	1.54	1.62	1.77	0.64	1.26	**2.34**			1.70	**2.17**	1.54		1.10		**2.26**	**2.53**
RUDN	1.12	1.08	1.32	**2.27**	**2.82**				1.26	**3.04**	1.65	1.35	0.62		**2.43**	**2.66**
BMSTU	1.87	1.94	**2.41**	0.98	1.67		1.70	1.26		0.52	**2.01**	1.84	1.64	**2.25**	1.83	1.89
KFU	1.97	**2.66**	**2.32**	0.82	**3.23**	**2.26**	**2.17**	**3.04**	0.52			**2.35**	0.87		**2.32**	**2.53**
TSU	**2.14**		**2.08**	0.23		1.46	1.54	1.65	**2.01**				1.16	1.83	**2.03**	1.81
1st MSMU	1.61	**4.65**	1.28	1.38	1.28			1.35	1.84	**2.35**			1.71			1.40
HSE	1.49	1.91	1.43	**2.55**	0.99	1.26	1.10	0.62	1.64	0.87	1.16	1.71			1.74	1.93
TPU	**2.09**		1.26			**2.08**			**2.25**		1.83				1.68	
MEPhI	**2.33**	**2.43**	**2.47**	1.24	**3.38**	1.79	**2.26**	**2.43**	1.83	**2.32**	**2.03**		1.74	1.68		1.77
MIPT	**2.16**	**2.71**	**2.42**			**2.10**	**2.53**	**2.54**	1.89	**2.53**	1.81	1.40	1.93		1.77	

There are only 13 out of 120 collaborative pairs in Table 1 for which ratio of Jaccard indices is less than one (these are shown as light blue cells). This means that for 89% pairs of universities their collaboration inside RISC Core is more important than in all other journals in RISC. Extramural collaboration plays a greater role in high-level research. A similar behaviour of Jaccard indices was found by Pislyakov and Shukshina (2014) for international collaboration of Russian organizations in producing highly cited papers.

Most remarkable are four pairs shown in Table 1 as yellow cells. These are Plekhanov University of Economics (REA) in collaboration with Kazan Federal University (KFU) and Moscow Engineering Physics Institute (MEPhI); Peoples' Friendship University (RUDN) again with Kazan Federal University (KFU); and the highest value in Table 1 is for the pair Financial University (FU)/Moscow Medical University (1st MSMU). In the latter case

collaboration of these institutions in the core is 4.65 times more important for them than in the whole RISC. They do not collaborate actively, but when they do so, most of their joint papers become published in the core (15 out of 20).

There are five organizations which become more important for any partner when the observation moves to the core publications. These are St. Petersburg University, Moscow Medical University, Tomsk Polytechnic University, Moscow Engineering Physics Institute and Moscow Institute of Physics and Technology. The latter, MIPT, also has the highest *average* ratio of Jaccards in Table 1 (2.16). And conversely (as we speak of bi-directional indicator), for these universities each other collaborator is also more important in the RISC Core than in the RISC as a whole database.

On the other side one can notice Academy of National Economy (RANEPA) performance. There are five cases (out of 13) of weakened collaboration when moving to the Core. Collaboration is not so important for RANEPA when preparing papers for the leading journals. Or, maybe, this relative negligence of collaboration in the RISC Core is a cause of the most unbalanced position of RANEPA concerning inside/outside Core publications observed in Figure 1.

Asymmetrical indexes: collaborative gain

More profound analysis of the patterns of inter-university collaboration may be obtained by investigating asymmetrical, unidirectional indexes which show how profitable is the collaboration for one partner and for another. Here we use an indicator of 'collaborative gain'. It is a ratio between percentage of papers found in the RISC Core for joint papers written by two institutions and share of Core papers in total output by one of them. In calculating the collaborative gain the numerator of the index is the same for both partner universities, it is a share of their collaborative papers which have entered the Core. But the denominator is different as soon as it is a share of all Core papers for a given institution. That is why A-B index is different from B-A.

Table 2. Collaborative gain received by university in a row from university in a column. (light blue cell: <1; empty cell: insufficient data)

	MSU	FU	SPBU	RANEPA	REA	UrFU	SFU	RUDN	BMSTU	KFU	TSU	1st MSMU	HSE	TPU	MEPhI	MIPT
MSU		0.48	1.72	0.69	0.55	1.97	1.25	0.90	1.59	1.77	2.29	1.64	1.44	2.00	2.36	2.36
FU	3.82		5.02	1.05	2.06	2.10	3.03	2.00	4.07	6.64		15.71	5.61		7.27	8.20
SPBU	1.90	0.70		0.60	0.65	1.83	1.43	1.05	2.06	2.12	2.34	1.37	1.44	1.24	2.63	2.80
RANEPA	5.38	1.03	4.19		1.38	1.78	1.21	4.23	2.08	2.09	0.89	4.77	7.60		3.79	
REA	3.82	1.80	4.05	1.23		2.19	2.35	5.17	3.53	8.18		4.47	3.02		10.64	
UrFU	2.25	0.30	1.89	0.26	0.36		1.74			2.03	1.75		1.28	2.05	1.99	2.63
SFU	2.92	0.89	3.02	0.36	0.79	3.56			1.89	2.78	2.85		1.66		3.63	4.49
RUDN	2.17	0.61	2.28	1.31	1.80				1.42	3.95	3.10	2.29	0.94		3.95	4.73
BMSTU	2.99	0.96	3.50	0.50	0.96		1.52	1.11		0.59	3.30	2.74	2.19	2.89	2.64	3.06
KFU	2.59	1.22	2.80	0.39	1.73	2.59	1.74	2.39	0.46			2.99	1.00		2.93	3.61
TSU	1.80		1.66	0.09		1.20	0.96	1.01	1.38				0.99	1.47	1.93	2.02
1st MSMU	1.51	1.82	1.13	0.56	0.59			0.87	1.34	1.88			1.57			1.67
HSE	1.58	0.77	1.42	1.07	0.48	1.23	0.78	0.42	1.27	0.75	1.38	1.87			1.92	2.46
TPU	2.35		1.31			2.11			1.80		2.20				1.91	
MEPhI	2.07	0.80	2.08	0.43	1.35	1.52	1.36	1.43	1.23	1.76	2.15		1.54	1.43		2.09
MIPT	1.30	0.56	1.38			1.26	1.05	1.02	0.89	1.35	1.41	0.99	1.23		1.31	
average	**2.56**	**0.92**	**2.50**	**0.66**	**1.06**	**1.94**	**1.53**	**1.80**	**1.79**	**2.76**	**2.15**	**3.88**	**2.25**	**1.85**	**3.49**	**3.34**

Collaborative gain indicator for 16 universities we study are summarized in Table 2. Collaborative gain may easily be interpreted in such a way: how more often university A

publishes its research in 'the best journals' if it collaborates with the university B? For example, value 3.82 in the row 3, column 2 cell of Table 2 means that collaborative papers by Financial University (FU) and Moscow University (MSU) have a 3.82 times higher share in RISC Core than all papers by FU.

Columns in Table 2 represent the 'gain', 'profit' which an organization brings to its partner. Cells marked by light blue color contain values less than one, this means that the gain is negative. In a sense this means that collaboration with this university (in a column) hinders its partner university (in a row). The probability of publishing such collaborative research in high-quality journals is lower than average for organization in a row.

We may notice that six out of 16 universities are *always* positive partners (see their columns in Table 2). Each other organization receives profit from making collaborative research with them. This is, of course, a natural situation as soon as collaboration should bring success. However, several institutions often bring difficulties with their partnership. These 'negative leaders' are Financial University (FU) and Academy of National Economy (RANEPA) with nine 'negative' cells in their columns, Plekhanov University of Economics (REA) following them with seven negative cases. Note that all three universities are among weakest in Figure 1. And they all are focused more on social sciences. This phenomenon should be scrutinized in further research.

The last row of Table 2 contains averages for all indexes for a given university in a column. Although these averages are somewhat artificial, still they may provide some insight into the overall value of the organization as a partner. The data from this row of Table 2 are sorted and organized into Figure 2.

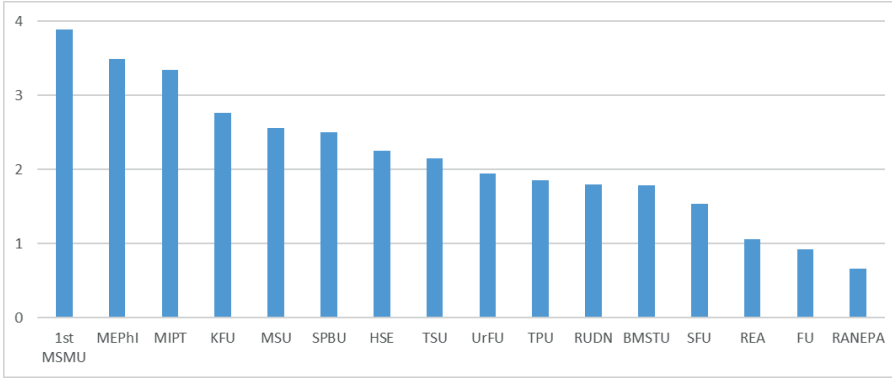


Figure 2. Average collaborative gain a university brings to its partners.

Finally, Table 2 also shows that there are three organizations which collaborations always bring gain *to themselves*. They are Financial University (FU), Plekhanov University of Economics (REA) and Tomsk Polytechnical University (TPU), their rows do not contain blue cells. The profitable strategy for these institutions is further strengthening of extramural collaborations.

Reciprocity of collaboration

Generally, collaboration is a win-win process when each of the partners receives some surplus compared to its sole performance. We observe such situation for the majority of collaborative

pairs in Table 2 when both A-B and B-A cells contain values greater than one and in this sense their partnership is *reciprocal*, profitable for both sides.

However, there are some remarkable exceptions. There are three pairs when collaboration is not very successful for both sides: Bauman University (BMSTU) with Kazan University (KFU); Tomsk State University (TSU) with Academy of National Economy (RANEPA); and the Higher School of Economics (HSE) with Peoples' Friendship University (RUDN). In each case there is negative collaborative effect for both institutions.

More often another situation may be found—when one of the partners gains from collaboration while another side loses. There are 9 cases of such kind of 'collaborative vampirism' for Financial University (FU), 8 for Academy of National Economy (RANEPA), 7 for Plekhanov University of Economics (REA), 2 for both Southern Federal University (SFU) and Peoples' Friendship University (RUDN), one for Bauman University (BMSTU), Kazan Federal University (KFU), Moscow Medical University (1st MSMU) and the Higher School of Economics (HSE). These are the cases of *non-reciprocal* inter-university collaboration, win-loss situation, when partnership benefits one side while harms another.

Still, even in win-win situations the gain one university brings to another is generally not equal to the gain it receives in return. To assess this asymmetry, we created Table 3 where differences of collaborative gains received and brought are calculated for each partnership. Positive values in Table 3 mean that university in a column has brought more gain to university in a row than it received in exchange. By definition, this table is skew-symmetric.

Table 3. Non-reciprocity (difference in collaborative gain).
How university in a column is *non-reciprocal* to university in a row.
(blue cell: < –3; light blue cell: <–0.1; yellow cell: between –0.1 and 0.1; empty cell: insuff. data)

	MSU	FU	SPBU	RANEPA	REA	UrFU	SFU	RUDN	BMSTU	KFU	TSU	1st MSMU	HSE	TPU	MEPhI	MIPT
MSU		-3.34	-0.19	-4.69	-3.27	-0.28	-1.67	-1.27	-1.40	-0.82	0.49	0.13	-0.14	-0.35	0.29	1.07
FU	3.34		4.32	0.02	0.27	1.79	2.14	1.40	3.11	5.42		13.90	4.84		6.47	7.64
SPBU	0.19	-4.32		-3.60	-3.40	-0.06	-1.59	-1.24	-1.44	-0.68	0.68	0.23	0.02	-0.07	0.55	1.42
RANEPA	4.69	-0.02	3.60		0.15	1.52	0.84	2.91	1.58	1.70	0.80	4.20	6.53		3.37	
REA	3.27	-0.27	3.40	-0.15		1.83	1.56	3.37	2.57	6.46		3.87	2.54		9.29	
UrFU	0.28	-1.79	0.06	-1.52	-1.83		-1.82			-0.56	0.55		0.05	-0.05	0.46	1.37
SFU	1.67	-2.14	1.59	-0.84	-1.56	1.82			0.37	1.04	1.89		0.88		2.27	3.44
RUDN	1.27	-1.40	1.24	-2.91	-3.37				0.31	1.56	2.09	1.42	0.51		2.52	3.71
BMSTU	1.40	-3.11	1.44	-1.58	-2.57		-0.37	-0.31		0.13	1.92	1.40	0.92	1.09	1.41	2.17
KFU	0.82	-5.42	0.68	-1.70	-6.46	0.56	-1.04	-1.56	-0.13			1.11	0.25		1.17	2.26
TSU	-0.49		-0.68	-0.80		-0.55	-1.89	-2.09	-1.92				-0.39	-0.73	-0.22	0.61
1st MSMU	-0.13	-13.90	-0.23	-4.20	-3.87			-1.42	-1.40	-1.11			-0.30			0.68
HSE	0.14	-4.84	-0.02	-6.53	-2.54	-0.05	-0.88	-0.51	-0.92	-0.25	0.39	0.30			0.38	1.23
TPU	0.35		0.07			0.05			-1.09		0.73				0.48	
MEPhI	-0.29	-6.47	-0.55	-3.37	-9.29	-0.46	-2.27	-2.52	-1.41	-1.17	0.22		-0.38	-0.48		0.79
MIPT	-1.07	-7.64	-1.42			-1.37	-3.44	-3.71	-2.17	-2.26	-0.61	-0.68	-1.23		-0.79	
average	**1.03**	**-4.20**	**0.89**	**-2.45**	**-3.14**	**0.40**	**-0.87**	**-0.58**	**-0.28**	**0.73**	**0.83**	**2.59**	**1.01**	**-0.10**	**1.97**	**2.20**

There are several pairs with almost perfect reciprocity, for example FU/RANEPA. Next, SPBU, UrFU, HSE and TPU also collaborate almost perfectly reciprocally with each other. We should remind, however, that according to our definitions the reciprocal collaboration may exist only between organizations of equal strength, in terms of share of their papers in the RISC Core (when denominators in the formula for collaborative gain are equal).

For each partner MIPT brings more gain than receives from them. The opposite is the case of Financial University (FU). Again, overall relative non-reciprocity of a given institution may be estimated with the column average of Table 3. Results are shown in Figure 3. This graph does not necessarily follow the sorting order from the higher to lower share of university's documents in RISC (as may be suggested by mathematical definition of 'gain'), at least because there are empty cells in Table 3, not all collaborations are active.

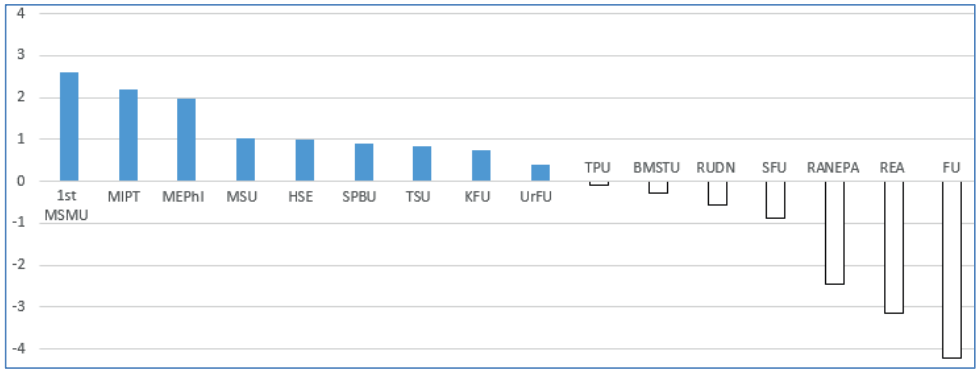


Figure 3. Average non-reciprocity in collaboration (how much a university brings to collaborations compared to how much it receives).

One may say that here we have nine 'donors', six 'acceptors' and one university (Tomsk Polytechnical) on the borderline. Of course, it is only an approximation and, additionally, made on the limited number of partner universities. Still it may shed light on the collaborative behavior and collaborative gains of the institutions we study. For example, the most significant change is for Kazan Federal University—from 4th place in Figure 2 to 8th in Figure 3. What does it mean? It means that KFU brings much gain in collaborations, but tends to collaborate mainly with those universities, who also give in return. As a result, it receives a lot too. Perhaps this is the most balanced and prudent though not altruistic approach to scientific collaboration.

Conclusion

We have found that study of collaboration between universities shows them in a more detailed manner, revealing their standing and roles in a complex network of collective scientific activity. We conclude that, in general, organizations strong by themselves are also strong collaborators. On the contrary, the universities with low share of papers published in the high-level journals are not very beneficial partners to the others. However, there are some institutions which break this pattern, for example one of the best collaborators appeared to be Moscow Medical University (1st MSMU), which has quite moderate position in Figure 1 showing performance of universities per se.

Our method of quality analysis by distribution of papers across journal subsets of different scientific level may be generalized. For example, those studies which analyze scattering of publications across journal quartiles use the same principle. They may focus only on the highest (first) quartile papers as in (Chowdhury & Philipson, 2016) or on full distribution by quartile (Olmeda-Gómez & Moya-Anegón, 2016). Quartiles may come from impact factor (Bordons & Barrigón, 1992), or SJR (Chinchilla-Rodríguez et al., 2015), or some other rankings. The idea of the method remains the same.

Moreover, this approach may be applied to Web of Science databases which have different status—namely Emerging Sources Citation Index as opposed to SCIE+SSCI+AHCI. As a small exercise, we have gathered statistics for different countries on how they publish their collaborative papers in different subsets of Web of Science. It appeared that, for example, collaboration of France and Germany is rather reciprocal and benefits both sides: joint Web of Science papers of these countries enter the 'main' journal indexes 2.8% and 2.6% more often than all publications of France and Germany, respectively. But at the same time when China partners with Iran, it publishes 5.1% more papers in 'emerging' index of WoS, while Iran gets 7.0% more its papers in the 'main' SCIE+SSCI+AHCI indexes, compared to its total WoS output. This is an example of non-reciprocal collaboration of countries measured with the international citation database.

As next steps of our research we plan to focus on disciplinary differences in collaboration and investigate cross-collaboration patterns between universities and Russian research institutes as observed from Russian Index of Science Citation and its Core.

References

Akoev, M., Moskaleva, O. & Pislyakov, V. (2018). Confidence and RISC: How Russian papers indexed in the national citation database Russian Index of Science Citation (RISC) characterize universities and research institutes. In *STI 2018 Conference Proceedings* (pp. 1328–1338). Leiden: Universiteit Leiden – CWTS. Retrieved January 8, 2019 from: https://openaccess.leidenuniv.nl/handle/1887/65344

Bordons, M. & Barrigón, S. (1992). Bibliometric analysis of publications of Spanish pharmacologists in the SCI (1984-89). Part II. Contribution to subfields other than "pharmacology & pharmacy" (ISI). *Scientometrics*, 25(3), 425–446. DOI: 10.1007/BF02016930

Chinchilla-Rodríguez, Z., Arencibia-Jorge, R., Moya-Anegón, F.D. & Corera Álvarez, E. (2015). Some patterns of Cuban scientific publication in Scopus: the current situation and challenges. *Scientometrics*, 103(3), 779–794. DOI:10.1007/s11192-015-1568-8

Chowdhury, G., Koya, K. & Philipson, P. (2016). Measuring the impact of research: Lessons from the UK's research excellence framework 2014. *PLoS ONE*, 11(6). DOI: 10.1371/journal.pone.0156978

Glänzel, W. (2001). National characteristics in international scientific co-authorship relations. *Scientometrics*, 51(1), 69–115. DOI: 10.1023/A:1010512628145

Glänzel, W. (2003). *Bibliometrics as a research field: A course on theory and application of bibliometric indicators*. Course Handouts, Leuven. Retrieved January 8, 2019 from: http://nsdl.niscair.res.in/jspui/bitstream/123456789/968/1/Bib_Module_KUL.pdf

Moskaleva, O., Pislyakov, V., Sterligov, I., Akoev, M. & Shabanova, S. (2018). Russian Index of Science Citation: Overview and review. *Scientometrics*, 116(1), 449–462. DOI: 10.1007/s11192-018-2758-y

Olmeda-Gómez, C. & Moya-Anegón, F.D. (2016). Publishing Trends in Library and Information Sciences across European Countries and Institutions. *Journal of Academic Librarianship*, 42(1), 27–37. DOI: 10.1016/j.acalib.2015.10.005

Pislyakov, V. & Shukshina, E. (2014). Measuring excellence in Russia: Highly cited papers, leading institutions, patterns of national and international collaboration. *Journal of the Association for Information Science and Technology*, 65(11), 2321–2330. DOI: 10.1002/asi.23093

Appendix 1. List of universities' abbreviations.

Abbreviation	Full name of university
1st MSMU	Sechenov First Moscow State Medical University
BMSTU	Bauman Moscow State Technical University

Abbreviation	Full name of university
FU	Financial University under the Government of the Russian Federation
HSE	National Research University Higher School of Economics
KFU	Kazan Volga Region Federal University
MEPhI	Moscow Engineering Physics Institute
MIPT	Moscow Institute of Physics and Technology
MSU	Moscow State University
RANEPA	Russian Presidential Academy of National Economy and Public Administration
REA	Plekhanov Russian University of Economics
RUDN	Peoples' Friendship University of Russia
SFU	Southern Federal University
SPBU	St. Petersburg State University
TPU	Tomsk Polytechnic University
TSU	Tomsk State University
UrFU	Ural Federal University

Appendix 2. Jaccard indices for inter-university collaboration in the entire RISC and RISC Core, % (empty cell: insufficient data).

	MSU		FU		SPBU		RANEPA		REA		UrFU		SFU		RUDN		BMSTU		KFU		TSU		1st MSMU		HSE		TPU		MEPhI		MIPT	
	RISC	Core	RISC	Core	RISC	Core	RISC	Core	RISC	Core	RISC	Core	RISC	Core	RISC	Core	RISC	Core	RISC	Core	RISC	Core	RISC	Core	RISC	Core	RISC	Core	RISC	Core	RISC	Core
MSU			0.19	0.15	0.38	0.69	0.26	0.29	0.19	0.15	0.18	0.36	0.16	0.25	0.35	0.39	0.71	1.32	0.25	0.49	0.53	1.14	0.35	0.56	0.72	1.07	0.07	0.16	1.41	3.29	1.30	2.82
FU	0.19	0.15			0.07	0.09	0.66	0.68	1.43	2.81	0.04	0.03	0.08	0.13	0.22	0.23	0.14	0.27	0.04	0.11			0.02	0.10	0.24	0.45			0.11	0.27	0.03	0.08
SPBU	0.38	0.69	0.07	0.09			0.21	0.21	0.02	0.02	0.14	0.27	0.05	0.09	0.13	0.16	0.08	0.20	0.10	0.24	0.16	0.33	0.08	0.11	0.28	0.40	0.04	0.05	0.23	0.58	0.10	0.23
RANEPA	0.26	0.29	0.66	0.68	0.21	0.21			0.97	1.27	0.12	0.06	0.15	0.10	0.29	0.65	0.07	0.07	0.05	0.04	0.03	0.01	0.09	0.13	0.35	0.90			0.03	0.04		
REA	0.19	0.15	1.43	2.81	0.02	0.02	0.97	1.27			0.03	0.02	0.07	0.09	0.22	0.62	0.19	0.31	0.09	0.31			0.06	0.08	0.18	0.18			0.07	0.23		
UrFU	0.18	0.36	0.04	0.03	0.14	0.27	0.12	0.06	0.03	0.02			0.05	0.12					0.06	0.13	0.15	0.22			0.09	0.11	0.06	0.12	0.08	0.15	0.08	0.17
SFU	0.16	0.25	0.08	0.13	0.05	0.09	0.15	0.10	0.07	0.09	0.05	0.12					0.04	0.06	0.04	0.09	0.08	0.13			0.04	0.04			0.07	0.15	0.07	0.19
RUDN	0.35	0.39	0.22	0.23	0.13	0.16	0.29	0.65	0.22	0.62							0.24	0.30	0.04	0.12	0.06	0.09	0.78	1.06	0.14	0.08			0.11	0.26	0.08	0.22
BMSTU	0.71	1.32	0.14	0.27	0.08	0.20	0.07	0.07	0.19	0.31			0.04	0.06	0.24	0.30			0.04	0.02	0.07	0.14	0.10	0.19	0.18	0.29	0.04	0.08	0.77	1.42	0.35	0.66
KFU	0.25	0.49	0.04	0.11	0.10	0.24	0.05	0.04	0.09	0.31	0.06	0.13	0.04	0.09	0.04	0.12	0.04	0.02					0.08	0.20	0.07	0.06			0.04	0.09	0.18	0.46
TSU	0.53	1.14			0.16	0.33	0.03	0.01			0.15	0.22	0.08	0.13	0.06	0.09	0.07	0.14							0.04	0.05	3.95	7.22	0.81	1.65	0.83	1.49
1st MSMU	0.35	0.56	0.02	0.10	0.08	0.11	0.09	0.13	0.06	0.08					0.78	1.06	0.10	0.19	0.08	0.20					0.03	0.06					0.14	0.19
HSE	0.72	1.07	0.24	0.45	0.28	0.40	0.35	0.90	0.18	0.18	0.09	0.11	0.04	0.04	0.14	0.08	0.18	0.29	0.07	0.06	0.04	0.05	0.03	0.06					0.13	0.22	0.76	1.46
TPU	0.07	0.16			0.04	0.05					0.06	0.12					0.04	0.08			3.95	7.22							0.28	0.47		
MEPhI	1.41	3.29	0.11	0.27	0.23	0.58	0.03	0.04	0.07	0.23	0.08	0.15	0.07	0.15	0.11	0.26	0.77	1.42	0.04	0.09	0.81	1.65			0.13	0.22	0.28	0.47			2.92	5.16
MIPT	1.30	2.82	0.03	0.08	0.10	0.23					0.08	0.17	0.07	0.19	0.09	0.22	0.35	0.66	0.18	0.46	0.83	1.49	0.14	0.19	0.76	1.46			2.92	5.16		

The impact of CSC Scholarships on scientific outputs and collaboration

Xuelian Pan [1] and Weina Hua[2]

[1] *xuelianpan@nju.edu.cn*
Nanjing University, School of Information Management, Nanjing, 210093 (P.R. China)

[2] *huawnl@nju.edu.cn*
Nanjing University, School of Information Management, Nanjing, 210093 (P.R. China)

Abstract

The China Scholarship Council (CSC) has provided financial assistance to Chinese citizens to research or study abroad and foreign citizens to study in China over twenty year, but little is known of the impact of CSC Scholarships. To fill this gap, this research-in-progress paper investigates the scientific outputs of CSC Scholarships using the data from Web of Science database. Moreover, we use several measures to describe the collaboration patterns of researchers supported by CSC Scholarships. Results show that the number of WoS publications supported by CSC Scholarships has dramatically increased during the period 2008-2017. We also find that CSC Scholarships promote Chinese researchers to collaborate with researchers from other countries/regions. Meanwhile, our results indicate that collaborating with authors from other countries could increase the international visibility and impact of papers of China.

Conference Topic

Scientific-scholarly internationalization, collaboration and mobility

Introduction

There is a general agreement that investments in science and technology have a significant impact on economic growth and social well-being (Lane & Bertuzzi, 2011). Many countries have substantially increased their research and development (R&D) expenditures to foster innovation and promote economic growth during the past decades. Global R&D expenditures have increased from 553 billion PPP$ (purchasing power parity dollars) in 1996 to 1,942 billion PPP$ in 2017 (UNESCO, 2018). As a developing country with fast-growing economy, China has increased its R&D expenditures from 14 billion PPP$ in 1996 to 451 billion PPP$ in 2016, with an annual growth rate of 18.96%. China has become a global leader in R&D investment; its gross domestic expenditure on R&D (GERD) in 2016 is only less than that of the United States. With these vast investments, Chinese government has provided various research funds and established numerous science scholarships to improve its scientific impact. The China Scholarship Council (CSC), a non-profit institution affiliated with the Chinese Ministry of Education, was established in 1996 by the Chinese government and it provides financial aids to Chinese citizens wishing to study abroad and to foreigners wishing to study in China (CSC, 2018). More than 250 thousand Chinese citizens including scholars, post-doctoral researchers, doctoral and master's students, and undergraduate students have been supported by CSC Scholarships to study or research abroad for months (ranging from 3 to 48 months) during the period 1996-2017. Meanwhile, CSC Scholarships (also called Chinese Government Scholarships) have also been provided to more than 350 thousand citizens of other countries/territories in the period 2005-2017 (China Education Yearbook, 2015).

There is a need for governments to verify that their investments in R&D benefit scientific knowledge production and innovation (Yan & Pan, 2015) and inform the general public about the impact of the investments (Lane, 2009). Historically, very limited resources have been

devoted to evaluations of public R&D investments (Lane & Bertuzzi, 2011). In recent years, increasing attention has been paid to analyse and assess the impact of government-sponsored R&D funding. Specifically, many studies on investigating the impact of research funding on scientific outputs have been conducted since some online databases such as Web of Science began to provide funding-acknowledgement data for their indexed publications (Shapira & Wang, 2010; Wang, Liu, Ding, & Wang, 2012; Yan, Wu, & Song, 2018). As yet, however, few studies have examined the impact of CSC Scholarships on scientific outputs.

Collaboration in research has been widely recognized as "a good thing" (He, Geng, & Campbell-Hunt, 2009). Numerous studies have explored the merits of research collaboration (Gazni & Didegah, 2011; Wuchty, Jones, & Uzzi, 2007). Many research institutions have launched initiatives to encourage their scientists to collaborate in scientific research because of the benefits of research collaboration (e.g., sharing knowledge and instruments, expediting research progress, and increasing the visibility of scientific outputs) (Gazni, Sugimoto, & Didegah, 2012). Moreover, scholars have examined the factors that influence research collaboration. Some main determinants of collaboration in scientific research such as funding policies (Ajiferuke, 2005), geographical distance (Plotnikova & Rake, 2014), and culture (Glänzel & Schubert, 2001) have been found. It seems reasonable to suppose that the CSC scholarships would promote international research collaboration by supporting researchers to go abroad to find international collaborators and reducing the geographical distance between international collaborators. However, this hypothesis has not been comparatively studied yet.

In this research-in-progress paper, we will not only investigate the scientific outputs of CSC scholarships, but also explore the collaboration patterns of the researchers supported by CSC scholarships. The results of this study will help funding agencies and scholars obtain a better understanding of the impact of funding or scholarships that provide financial assistance to citizens wishing to research or study abroad.

Methods

The Web of Science (WoS) database was selected as data source for publications supported by CSC scholarships because most universities and research institutions in China value publications published in the journals that indexed in WoS and offer rewards to encourage their students and researchers to publish their papers in WoS journals (Quan, Chen, & Shu, 2017). "China Scholarship* Council", "Chinese Scholarship* Council", "CSC Scholarship*", "Chinese Government Scholarship*", and "China Government Scholarship*" were used as search terms. The timespan was limited to 2008-2017 and document types were limited to article or review. We first searched WoS (limited to SCIE, SSCI and A&HCI) for all papers containing our search terms in the "Funding" field in December 20, 2018. The bibliographic information of these papers was downloaded and then imported into SQL Server 2012 (Microsoft, Redmond, WA, USA) for further analysis. In total, we obtained 41,622 papers supported by CSC scholarships. We then classified these papers into three groups according to the type of collaboration as follows:

(1) Non-collaborative papers (solo-authored papers), including non-collaborative papers of China and non-collaborative papers of other counties;
(2) National collaborative papers (co-authored papers with all the authors from one country), including Chinese domestic collaborative papers (co-authored papers with all the authors from China) and other countries' national collaborative papers (co-authored papers with all the authors from a country other than China);
(3) International collaborative papers (co-authored papers with more than one country address), including international collaborative papers of China (co-authored papers with at least one author affiliated to a Chinese institution and one or more authors affiliated to foreign institutions) and international collaborative papers of other countries (co-

authored papers with at least two authors from different countries and with no authors from China).

Next, we used several measures, such as percentage of collaborative papers, percentage of international collaborative papers, and average citations per paper, to describe the collaboration patterns of researchers supported by CSC Scholarships.

Preliminary results

Scientific publications supported by CSC Scholarships

As shown in Figure 1, there has been a rapid and sustained increase in the number of WoS papers supported by CSC Scholarships during the period 2008-2017. The number of such papers has increased from 184 in 2008 to 10,434 in 2017, with an annual growth rate of 57%. This rapid growth may relate to the increase of the number of citizens supported by CSC Scholarships in the same period. Both the number of Chinese citizens and the number of citizens of other countries/territories supported by CSC Scholarships have increased steadily over the past decade: the number of Chinese citizens has increased from 12,957 in 2008 to 32,500 in 2017; the number of citizens of other countries/territories has increased from 13,516 in 2008 to 58,600 in 2017.

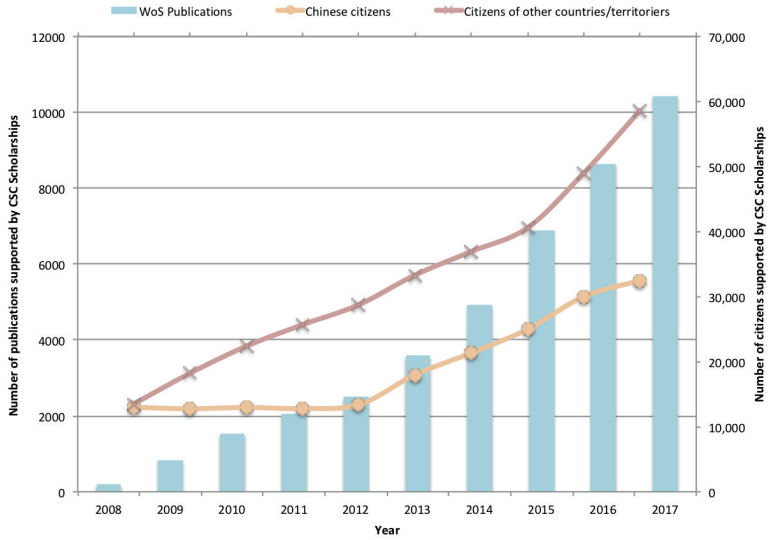


Figure 1. Scientific publications and citizens supported by CSC Scholarships by year.

Our dataset covers 143 of all the 151 WoS research areas. Chemistry leads with 8,679 papers and a share of 21% of total papers supported by CSC scholarships. It is followed by Engineering (8,461, 20%), Materials Science (7,672, 18%), and Physics (5,489, 13%). The top 47 categories with the greatest number of papers belong to natural science. Business & Economics, ranked 48th in the total ranking, leads in the social sciences and arts humanities categories with 242 papers. The papers were further grouped into five broad categories (i.e., Arts & Humanities, Life Sciences & Biomedicine, Physical Sciences, Social Sciences, and Technology) based on their research areas and a list of WoS research areas with their broad categories provided in WoS website (https://images.webofknowledge.com/images/help/WOS

/hp_research_areas_easca.html). Technology leads with 20,833 (50.05% of the total 41,622) papers supported by CSC Scholarships, followed by Physical Sciences (20,036, 48.14%), Life Sciences & Biomedicine (11,906, 28.61%), Social Sciences (581, 1.40%), and Arts & Humanities (25, 0.06%).

Collaboration patterns of researchers supported by CSC Scholarships

Among the total of 41,622 papers supported by CSC Scholarships, 99% of the papers are co-authored and 68% of the papers are internationally co-authored. The percentage of national collaborative papers remains stable over the period analysis, while the percentages of international collaborative papers in the period 2008-2012 (which range from 72% to 86%) are higher than those in the period 2013-2017 (which range from 65% to 68%). Moreover, some differences were found among the research areas in the percentage of international collaborative papers. Of the 70 research areas with more than 100 papers, Astronomy & Astrophysics and Entomology lead with 83% and 82% of international collaborative papers, respectively. However, Mathematics (58%), Fisheries (58%), and Toxicology (59%) have a lower percentage of international collaborative papers than the others. These figures are higher than those found in Gazni et al.'s (2012) study of WoS papers published in 2009, which reported the highest such percentage of 49%.

From the total 41,622 papers supported by CSC Scholarships, we identified 29,941 papers with at least one author affiliated to Chinese institutions. Among these China's papers, 84% were international collaborative papers. These papers were then classified into different groups according to their research areas and the percentage of international collaborative papers for each group with more than 100 papers was calculated. Some research areas, such as Evolutionary Biology (96.30%) and Hematology (95.52%), had a higher percentage of international collaborative papers than other research areas. In contrast, Mathematics (62.03%) had a lower such percentage than the others. Among the five broad categories, Arts & Humanities (63%) and Social Sciences (79%) have a lower percentage of international collaborative papers than Physical Sciences (83%), Technology (85%), and Life Sciences & Biomedicine (88%). Compared to the percentage of international collaborative papers in China's SCI and SSCI publications (which reported 23% and 46%, respectively) found in a previous study (Liu, Hu, Tang, & Wang, 2015), that of China in WoS papers supported by CSC Scholarships is higher. These findings demonstrates that our hypothesis for this study is correct; CSC scholarships promote international research collaboration by supporting researchers to go abroad to find international collaborators and reducing the geographical distance between international collaborators.

Chinese researchers collaborated with authors from 122 countries in the 25,237 papers supported by CSC Scholarship. The United States was the top collaborator (12,332, 48.86%) of China among all the collaborating countries, followed by Australia (2,377, 9.42%), Canada (2,266, 8.98%), England (2,109, 8.36%), and Germany (1,964, 7.78%). Of the top 20 collaborators of China, only Pakistan is a developing country. These findings suggest that Chinese researchers are more likely to collaborate with researchers from developed countries and provide evidence that United States is China's most important partner, which is consistent with the observations of previous studies on China's SSCI papers (Liu, Hu, Tang, & Wang, 2015) and China's SSCI and A&HCI publications (Li & Li, 2015). Moreover, among these China's major collaborators, four countries (i.e., Japan, Singapore, South Korea, and Pakistan) are in Asia, which constitute a small fraction of the international collaborative papers of China supported by CSC Scholarship.

Table 1 presents the citation counts of the non-collaborative, national collaborative, and international collaborative papers. We can find that the 416 non-collaborative papers received a total of 1,982 citations, with an average of 4.76 citations per paper and a median of 2. Among

the non-collaborative papers, the China's papers have higher average citations per paper than other countries' papers. Compared to the non-collaborative papers, the national collaborative papers received more citations on average (14.04 citations per paper) over the study period. However, on average, they received fewer citations per paper than the international collaborative papers (17.63 citations per paper). The results suggest that collaborative papers receive more citations than non-collaborative papers, which is consistent with the findings of previous studies (Li & Li, 2015; Wuchty, Jones, & Uzzi, 2007; Beaver, 2004). Moreover, an interesting finding is that the mean citations of the collaborative papers with all the authors from other countries/regions (17.28) is 2.2 times as large as that of the collaborative papers with all the authors from China (7.80). Another interesting finding is that the mean citations of the international collaborative papers with at least one author affiliated to Chinese institutions (17.72) is larger than that of all other subgroups of papers. These findings suggest that Chinese researchers may be able to improve the impact of their publications by collaborating with researchers from other counties.

Table 1. Citation counts of the non-collaborative and collaborative papers.

Group	*Subgroup*	*Publications*	*Total citations*	*Mean citations*	*Median citations*
Non-collaborative papers	China	325	1,325	4.08	2
	Other countries	91	657	7.22	3
	Total	416	1,982	4.76	2
National collaborative papers	China	4,379	34,167	7.80	3
	Other countries	8,425	145,607	17.28	8
	Total	12,804	179,774	14.04	6
International collaborative papers	China	25,237	447,088	17.72	8
	Other countries	3,165	53,689	16.96	9
	Total	28,402	500,777	17.63	8

Conclusion and future work

The CSC Scholarships, as a government investment, have been offered to more than 500 thousand people during the past decades. However, we still know little about the scientific outputs and the effectiveness of this investment. Therefore, there is a need to assess its impact on scientific outputs and international collaboration. In this research-in-progress paper, we have investigated the scientific outputs of CSC Scholarships using WoS database. Results show that the number of WoS papers supported by CSC Scholarships has dramatically increased over the past decade. Moreover, we have examined the collaboration patterns of the researchers financially supported by CSC Scholarships. It is not too surprising to find that CSC Scholarships have promoted Chinese researchers to collaborate with researchers from other countries/regions. Our results also show that Chinese authors would be able to increase the international visibility and impact of their papers by collaborating with authors from other countries.

A number of interesting research questions still remain to be investigated. For example, what types of CSC Scholarships have higher scientific productivity? Whether publications supported by CSC Scholarships receive more citations than those supported by other Chinese funding sources? What is the relationship between CSC Scholarships and citation impact? Whether CSC Scholarships are more likely to promote international collaboration than other Chinese funding sources? What is the impact of CSC scholarships on the subsequent career outcomes of the

doctoral students supported? The answers to the above questions will provide us a better understanding of the impact of CSC Scholarships and help policy makers and administrators make a better decision on effective use of public investment.

Acknowledgments

This work was funded by the National Natural Science Foundation of China (Grant No. 71704077).

References

Ajiferuke, I. (2005). Inter-university collaboration in Canada. *The Canadian Journal of Information and Library Science*, *29*(4), 407–418.

Beaver, D. D. (2004). Does collaborative research have greater epistemic authority?. *Scientometrics*, *60*(3), 399-408.

China Education Yearbook Editorial Board. (2015). *China education yearbook 2015*. Beijing: People's Education Press.

CSC. (2018). China Scholarship Council Official Website. Retrieved from https://www.csc.edu.cn/

Gazni, A., & Didegah, F. (2011). Investigating different types of research collaboration and citation impact: A case study of Harvard University's publications. *Scientometrics*, *87*(2), 251–265.

Gazni, A., Sugimoto, C. R., & Didegah, F. (2012). Mapping world scientific collaboration: Authors, institutions, and countries. *Journal of the American Society for Information Science and Technology*, *63*(2), 323-335.

Glänzel, W., & Schubert, A. (2001). Double effort = double impact? A critical view at international co-authorship in chemistry. *Scientometrics*, *50*(2), 199-214.

He, Z. L., Geng, X. S., & Campbell-Hunt, C. (2009). Research collaboration and research output: A longitudinal study of 65 biomedical scientists in a New Zealand university. *Research Policy*, *38*(2), 306-317.

Lane, J. (2009). Assessing the impact of science funding. *Science*, *324(*5932), 1273-1275.

Lane, J., & Bertuzzi, S. (2011). Measuring the results of science investments. *Science, 331*(6018), 678-680.

Li, J., & Li, Y. (2015). Patterns and evolution of coauthorship in China's humanities and social sciences. *Scientometrics*, *102*(3), 1997-2010.

Liu, W., Hu, G., Tang, L., & Wang, Y. (2015). China's global growth in social science research: Uncovering evidence from bibliometric analyses of SSCI publications (1978–2013). *Journal of Informetrics*, *9*(3), 555-569.

Plotnikova, T., & Rake, B. (2014). Collaboration in pharmaceutical research: Exploration of country-level determinants. *Scientometrics*, *98*(2), 1173-1202.

Quan, W., Chen, B., & Shu, F. (2017). Publish or impoverish: An investigation of the monetary reward system of science in China (1999-2016). *Aslib Journal of Information Management*, *69*(5), 486-502.

Shapira, P., & Wang, J. (2010). Follow the money. *Nature*, *468*(7324), 627.

UNESCO. (2018). Science, technology and innovation: Gross domestic expenditure on R&D (GERD), GERD as a percentage of GDP, GERD per capita and GERD per researcher. Retrieved from http://data.uis.unesco.org/?queryid=181#

Wang, X., Liu, D., Ding, K., & Wang, X. (2012). Science funding and research output: A study on 10 countries. *Scientometrics*, *91*(2), 591-599.

Wuchty, S., Jones, B. F., & Uzzi, B. (2007). The increasing dominance of teams in production of knowledge. *Science*, *316*(5827), 1036-1039.

Yan, E., & Pan, X. (2015). A Bootstrapping Method to Assess Software Impact in Full-Text Papers. In *ISSI 2015*.

Yan, E., Wu, C., & Song, M. (2018). The funding factor: A cross-disciplinary examination of the association between research funding and citation impact. *Scientometrics*, *115*(1), 369-384.

Medical research versus medical needs in Africa

Hugo Confraria[1] and Lili Wang[2]

[1] *h.confraria@sussex.ac.uk*
SPRU – Science Policy Research Unit, University of Sussex (United Kingdom)

[2] *wang@merit.unu.edu*
UNU-MERIT, Maastricht University (The Netherlands)

Abstract

Africa is a continent facing severe, urgent, and often unique health challenges. However, due to the deficiencies in financial support from local governments, medical research in Africa has been greatly dependent on foreign partners. Hence there is a rising concern that research conducted in African countries will follow topics determined by foreign funders for their purposes and consequently fail to respond to specific local health needs. In this article, we investigate whether the distribution of medical research priorities and investment in medical research, across diseases in Africa, is related to the medical/health needs of local. Different from the arguments in some existing literature, we find that high dependence on international donors is not associated with less alignment between local health needs and local medical research efforts. Our results show that in sub-Saharan Africa most health fields with a higher disease burden are also the ones with relatively more medical scientific publications. Among all the regions, Eastern Africa presents the highest positive association between disease burden and research efforts, and Northern Africa is the region where these two dimensions are less aligned. We also find that certain international research funders play a vital role in health research in Africa.

Introduction

Africa is a continent facing severe, urgent, and often unique health challenges. The region has made overall progress during the last decades in reducing mortality and prolonging life but its burden of disease per population continues to be two times higher than that of higher income countries. At the same time, most African countries have difficulties in supporting medical research, and the pharmaceutical industry may be reluctant to sponsor research in lower income countries because the prospects of profit are limited, even if effective treatments are developed (World Health Organization, 2012).

Nevertheless, it has been well recognized that health research conducted in low-income countries is of great importance. Medical research done and applied in lower income contexts can provide enormous contributions to discovering previously unknown diseases which have substantial social and economic impact on the world (Lee et al., 2013). Furthermore, medical research based in low-income countries can help researchers have a clear understanding of the local constraints and barriers to and facilitators of the implementation of research in practice.

Lack of funding for research is the major barrier to the development of clinical research capacity in Africa and the majority of clinical research is based on funding from external donors (Cardoso et al., 2014). Since local governments are taking a peripheral role (Cardoso et al., 2014), there is a risk that research conducted in African countries will follow topics determined by foreign funders for their purposes and consequently fail to respond to specific local health needs (Binka, 2005; Gaillard, 1994). There is a rising concern that there is lack of alignment between research efforts and health problems (Atal et al., 2018; Evans et al., 2014; Rafols & Yegros, 2017; Røttingen et al., 2013).

In our research, we will look specifically at the scientific output of African researchers and also at the funding institutions acknowledged in their publications. This will allow us to evaluate whether international development funders and pharmaceutical companies are supporting research that matches the research needs of African regions.

This approach and research question may be interesting for two main reasons. First, due to the tremendous health challenges the continent faces, improved Africa-relevant health research can have an important role in changing the professional practice of health care providers and a significant impact on health outcomes. Second, the improvement in research capabilities in the continent in the health sciences may demonstrate that persistent support and funding from development partners such as the Wellcome Trust, NIH, European Union or Bill and Melinda Gates Foundation, might pay off.

This study uses DALYs (Disability-Adjusted Life Years) in each disease field and African region as a proxy for societal needs in health, which is compared with scientific research in each corresponding disease field and region. We focus on four African regions (Eastern Africa, Northern Africa, Southern Africa or West & Central Africa) as defined by the UN classification, and our central research questions are: 1) Is the amount of research produced on various diseases by African researchers related to their countries' burden of disease? 2) What kind of health research is being funded by international organizations? 3) What are the main drivers of medical research in Africa?

Findings

Disease burden vs. research output

There are vast imbalances in global health between Africa and higher income regions. In 2015, the estimated world share of scientific output in medical and health-related areas by African researchers accounted to only 2.8%, a marked contrast to the fact that almost 26% of the global disease burden in 2015 was in Africa. Fig. 1 shows that, different from the worldwide average, the disease burden per capita is much higher than medical research output per capita in all African regions. This mismatch is especially remarkable for Eastern Africa and West & Central Africa.

It is well known that Africa's scientific output is highly skewed across nations and disciplinary areas, with South Africa and Egypt representing around 50% of total African output, and their specialization being mostly on agricultural sciences given the needs of the continent and in health-related sciences due to research work on tropical diseases and specific health problems, as well as from the location of international medical research centres on African soil, and the abundance of international cooperation between African researchers and those overseas (AOSTI 2014; Confraria & Godinho 2014; Tijssen 2007; UNESCO 2015). We find that medical research per capita has increased in all regions around 50%, between 2006-2010 and 2011-2015, and that the disease with more publications in each region is "parasitic and vector diseases" in Eastern and Western & Central Africa, "malignant neoplasms" in Northern Africa and "HIV/AIDS" in Southern Africa.

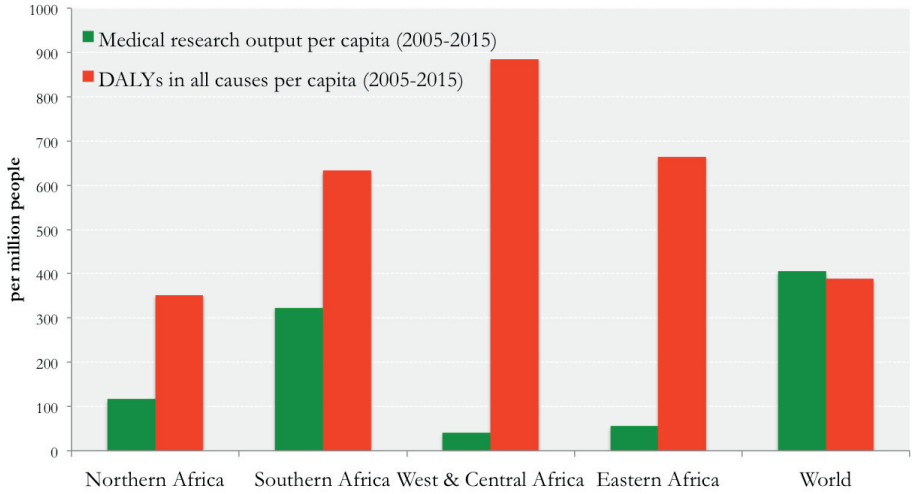


Fig. 1. African regions medical scientific production, and disease burden per capita.
Note: Own calculation based on Web of Science (WoS) and World Health Organization (WHO). In this chart, the numbers presented are yearly averages. For the medical research output, the calculation only takes to account the 28 diseases identified (Type I and II diseases). For the DALYs, the calculation is for all causes (Type I, II and III diseases).

As for DALYs we find that the countries with the highest incidence by African region are: Egypt (38% of total) for Northern Africa; South Africa (87% of total) for Southern Africa; Nigeria (40% of total) for West & Central Africa and Ethiopia (23% of total) for Eastern Africa. DALYs per capita are decreasing on average in all regions, but the most negative growth rates were in Eastern and Western & Central Africa.

Disease burden specialization vs. research specialization

To further assess the association between research output and disease burden, in Fig. 2 we display the disease burden specialization (relative to the world) as a function the medical research specialization (relative to the world) in 28 diseases (type I and II excluding type III diseases) by African region.

Eastern Africa and West & Central Africa exhibit a strong positive association between the two dimensions. The interpretation is that the diseases that have a higher disease burden relative to the World, are also the ones that have higher scientific specialization. On the other hand, this positive association is not so clear for Southern Africa and especially for Northern Africa. In these regions, there are some diseases like “parasitic and vector diseases” and “leprosy” that display a low level of disease burden specialization but a high level of scientific specialization. One could argue that these topics are “over-researched” since the disease burden in these regions is not relatively high. However, since in 2015 globally, 4 out of 5 DALYs in “parasitic and vector diseases” are from Africa, due to the high disease burden in the Eastern Africa and West & Central Africa, this high level of scientific specialization may be justified by the existence of research tradition in these areas, promotion of intra-African research collaborations and the development of research capacities in other African regions. As for leprosy, since it is considered a neglected disease, with residual impact in high-income countries, the existence of scientific capabilities in these regions may be justified.

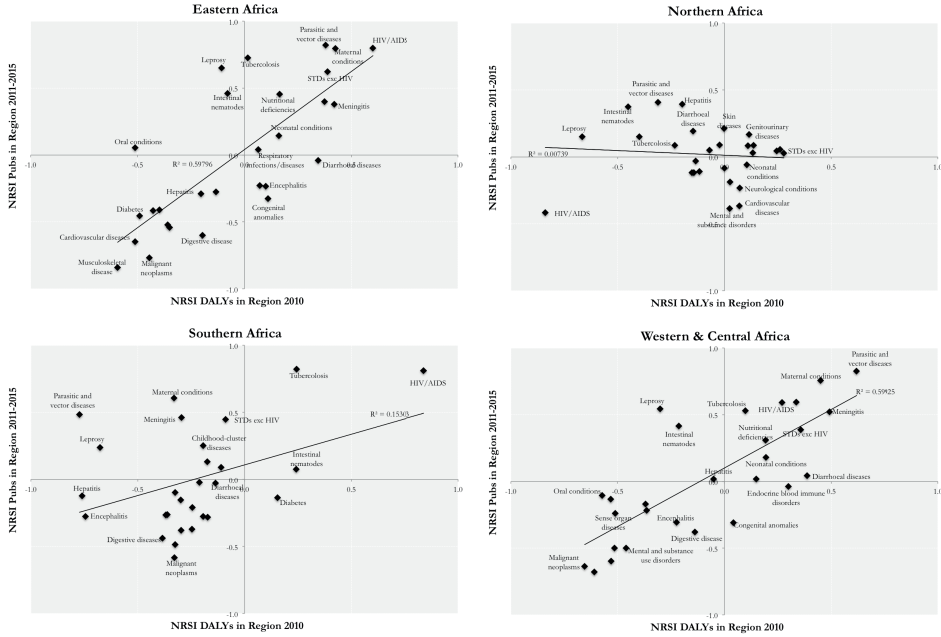


Fig. 2. Disease burden specialization vs research specialization by disease.

Note: Own calculation based on WoS and WHO. DALYs from 2010 and scientific publications from 2011-2015.

Another interesting finding is that in all African regions "tuberculosis" is a topic of high scientific specialization. Since the risk of developing tuberculosis is estimated to be between 16-27 times greater in people living with HIV than among those without HIV infection[1], the high levels of scientific specialization in "tuberculosis" may be because in most countries HIV/AIDS research is done in conjunction with tuberculosis research. In South Africa, for example, most patients who die from HIV-related causes die from tuberculosis or similar illnesses.

Finally, we do not find any region where a disease with relatively high burden (NSI DALYs > 0.5) is not a scientific priority for a region (NSI Pubs > 0.0). The only diseases that could be seen as "under-researched" are the ones that are between 0 and 0.5 in the x-axis (NSI DALYs) and between -0.5 and 0 in the y-axis (NSI Pubs) in every region. Some examples include "diabetes" in Southern Africa and "diarrheal diseases" in Eastern Africa.

Finally, it is important to note the high correlation coefficients between Eastern Africa specializations and West & Central Africa specializations. These two regions are highly scientifically specialized in diseases where they have a relatively high disease burden and have very few publications in diseases where they have a relatively low disease burden. Since scientific institutions in many of these African countries suffer from specific challenges such as poor conditions for research personnel, heavy teaching loads, inability to mentor young scholars, inadequate infrastructure and lack of funding (Mouton, 2008; Sawyerr, 2014), a potential reason for this match between disease burden and medical research may be due to efficient resource allocation or the influence of external funders (Ndounga Diakou et al., 2017).

[1] http://www.who.int/hiv/topics/tb/en/
[2] http://icd9.chrisendres.com/
[3] One of the reviewers is a PhD student in international health and development, who worked for five years as a nurse in epidemic contexts in

Funded Medical Research in Africa

The share of the total medical research funded by at least one funding institution in each African region, across the disease categories, ranged from 37% in Northern Africa to 79% in Eastern Africa (Fig. 3).

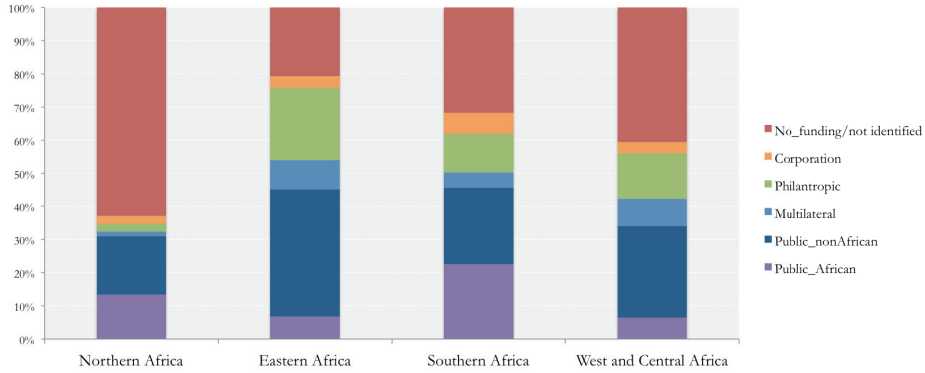


Fig. 3. Share of publications by funding type (2011-2015).
Note: Own calculation based on WoS.

The highest share of research funding in all regions is from public non-African funding institutions (e.g. NIH, EU, USAID, Medical Research Council (UK)), followed by Philanthropic funding institutions (e.g. Wellcome Trust, Gates Foundation) that make particularly relevant contributions in Eastern African countries. Public African funding institutions have higher shares of funding in Southern Africa (e.g. National Research Foundation (ZA), Medical Research Council (ZA)) and Northern Africa (e.g. Tunisian Government, Egyptian Government). In all regions the contribution of corporations is relatively small but pharmaceutical producers like GlaxoSmithKline, Pfizer and Novartis were acknowledged in 328, 302 and 238 publications between 2011 and 2015 respectively. Multilateral funding institutions like WHO, EDCTP and the World Bank are mostly funding medical research in Eastern African countries and West & Central African countries. While doing this analysis we find a significant overlap between public non-African and multilateral funding (50% of all publications with multilateral funding also acknowledge a public non-African funder). Since theoretically it is difficult to distinguish what are the different reasons that lead public non-African and multilateral institutions to fund African health research, further analysis in this article will consider the two categories as the same category (public non-African).

Research specialization of funders

We are also interested to know if each funder (or group of funders) supports research in specific diseases. Table 1, highlights the top 20 funders in Africa by disease between 2011 and 2015. One key finding is that "Parasitic and vector diseases", "HIV/AIDS" and "Tuberculosis" are a priority for every top10 funder. These results are in line with Chapman et al. (2017) that also find that three diseases – HIV/AIDS, malaria and tuberculosis – collectively received more than two-thirds ($2,247m, 70%) of all global funding for neglected disease R&D in 2016.

Table 1. Share of research funded by the top 20 funding institutions by disease

Disease	Total Pubs (2011-2015)	NIH_US	Wellcome_Trust	NRF_ZA	Gates_Foundation	EU	Fogarty_Int_Center	MRC_ZA	WHO	USAID	CDC_US	NIAID_US	MRC_UK	Tunisian_Gov	DFID_UK	EDCTP	GlaxoSmithKline	Univ_Cape_Town	DST_ZA	PEPFAR_US	Pfizer
Parasitic and vector diseases	7996	13%	11%	3%	9%	5%	2%	1%	4%	3%	2%	2%	2%	0%	2%	2%	1%	1%	1%	0%	0%
HIV/AIDS	4230	20%	7%	5%	5%	2%	5%	3%	2%	5%	4%	4%	2%	0%	1%	2%	1%	1%	1%	4%	1%
Tuberculosis	3402	21%	10%	7%	4%	5%	6%	5%	2%	6%	3%	3%	3%	0%	1%	5%	0%	2%	2%	3%	1%
Diarrhoeal diseases	3079	4%	2%	5%	4%	1%	1%	1%	2%	1%	1%	0%	0%	2%	1%	0%	1%	0%	1%	0%	0%
Malignant neoplasms	2721	6%	0%	4%	0%	1%	1%	1%	0%	0%	0%	0%	0%	2%	0%	0%	1%	1%	0%	0%	0%
Resp. infections/diseases	2183	6%	6%	4%	4%	2%	1%	2%	2%	2%	2%	1%	2%	1%	1%	0%	4%	1%	1%	0%	3%
Diabetes mellitus	2150	4%	2%	4%	1%	2%	1%	3%	1%	0%	0%	0%	1%	2%	1%	0%	1%	1%	0%	0%	1%
Endocrine blood immune disorders	2144	8%	5%	2%	2%	2%	2%	1%	0%	1%	1%	1%	1%	2%	1%	1%	0%	0%	0%	0%	1%
Neurological conditions	2000	6%	4%	4%	1%	2%	2%	2%	1%	0%	0%	0%	1%	1%	1%	0%	1%	2%	0%	0%	1%
Genitourinary diseases	1967	5%	1%	2%	1%	1%	1%	2%	1%	1%	0%	0%	1%	1%	0%	0%	0%	1%	0%	0%	1%
Cardiovascular diseases	1814	6%	2%	6%	0%	1%	2%	6%	1%	0%	0%	0%	1%	2%	0%	0%	1%	2%	0%	0%	2%
Mental and substance use disorders	1733	10%	3%	4%	1%	3%	3%	5%	2%	1%	0%	0%	1%	1%	1%	0%	3%	2%	1%	0%	4%
Digestive disease	1490	3%	1%	2%	1%	1%	1%	1%	2%	0%	0%	0%	0%	1%	0%	0%	1%	0%	0%	0%	0%
Hepatitis	1296	6%	2%	4%	1%	2%	1%	1%	0%	1%	0%	1%	1%	1%	0%	0%	1%	0%	0%	0%	0%
Musculoskeletal disease	1088	2%	2%	2%	1%	1%	1%	3%	1%	0%	0%	0%	1%	1%	0%	0%	1%	0%	0%	0%	3%
Intestinal nematode	1017	5%	7%	4%	3%	3%	1%	0%	2%	1%	1%	0%	2%	2%	1%	0%	1%	0%	0%	0%	0%
Sense organ diseases	888	5%	4%	1%	1%	1%	0%	1%	1%	0%	0%	0%	0%	2%	0%	0%	0%	0%	0%	0%	0%
Skin diseases	837	2%	3%	3%	1%	2%	0%	0%	1%	1%	0%	0%	0%	1%	1%	0%	0%	0%	0%	0%	0%
Meningitis	599	14%	11%	6%	7%	3%	6%	3%	4%	3%	2%	2%	3%	1%	1%	1%	2%	1%	1%	1%	3%
Congenital anomalies	470	4%	1%	1%	1%	1%	0%	1%	0%	0%	0%	0%	1%	0%	0%	0%	0%	0%	0%	0%	0%
STDs exc HIV	424	21%	8%	2%	9%	2%	2%	1%	2%	5%	4%	2%	3%	1%	4%	3%	1%	1%	0%	4%	0%
Neonatal conditions	392	9%	4%	1%	10%	2%	1%	2%	4%	4%	1%	1%	1%	0%	1%	1%	0%	0%	0%	0%	0%
Maternal conditions	380	3%	3%	1%	6%	1%	0%	1%	7%	4%	0%	0%	1%	0%	4%	0%	0%	1%	0%	2%	0%
Childhood-cluster diseases	354	6%	5%	3%	7%	10%	3%	1%	10%	2%	10%	1%	3%	0%	2%	0%	1%	0%	1%	1%	1%
Nutritional deficiencies	347	8%	7%	1%	4%	3%	1%	2%	2%	5%	2%	0%	2%	1%	1%	0%	1%	1%	0%	0%	0%
Oral conditions	311	5%	3%	2%	0%	4%	1%	1%	2%	2%	0%	1%	1%	0%	2%	0%	0%	0%	1%	0%	0%
Encephalitis	86	6%	3%	5%	0%	3%	2%	0%	2%	1%	2%	0%	1%	0%	0%	0%	0%	0%	0%	1%	0%
Leprosy	78	5%	1%	3%	1%	3%	1%	1%	4%	0%	0%	3%	0%	0%	1%	0%	1%	0%	0%	0%	0%

The only funders that are not so biased towards these three diseases and have a more horizontal distribution of their funding are the National Research Council (ZA), Medical Research Council (ZA) and Tunisian Government. These are all African funders that may have different priorities than international organizations. It is, however, important to notice the absence of Public African funding in Eastern Africa and West & Central Africa. In these regions research is heavily dependent upon foreign funds.

Interestingly, Gates Foundation funds more than 10% of African research on "neonatal conditions" which is the disease with the highest absolute disease burden in Eastern Africa and West & Central Africa. It has been argued that Gates Foundation investment has tried to balance the public sector focus on basic research (Chapman et al. 2017). According to G-finder data, it has provided 55% of all funding to neglected diseases in the world to product development partnerships and 47% of all funding for platform technologies between 2007 and 2016.

Overall, public non-African and philanthropic groups fund similar diseases, and in "Eastern Africa", "Southern Africa" and "West & Central Africa" they are mostly focused on medical research in "parasitic and vector diseases", "tuberculosis" and "HIV/AIDS". The share of total funding from philanthropic and public non-African institutions to "parasitic and vector diseases" is particularly high in "West & Central Africa" and "Eastern Africa". It represents more than 40% of the total funding of these institutions in both regions. "Parasitic and vector diseases" group includes diseases such as malaria, dengue, trachoma, yellow fever, rabies, chagas disease, among others. Malaria is by far the condition that leads to higher disease burden in this category. According to Head et al., (2017) global research funding for malaria

in sub-Saharan Africa is mostly allocated to Tanzania, Uganda, Kenya, Malawi, Ghana, and Nigeria. These are locations with a track record of success in similar projects where it is perceived that investments will make a positive difference and where any research will be feasible. The research supported by corporations is substantially higher in absolute terms in Southern Africa, and in areas such as diabetes, cardiovascular diseases and respiratory infections/diseases.

Econometric analysis

In this section, we present the results of the estimation of Equation (3) (see Data and Methods section). We pooled data from the period 2011-2015 for research specialization and 2010 for disease burden specialization. After constraining our database to diseases in regions with a minimum of 50 publications (to avoid outliers when computing the Normalized Specialization Index - NSI), we end up with 103 observations from our four African regions in the 28 diseases. Table 2 shows that the Eastern African region is the region where the association between disease burden and research specialization is the highest. The region is highly dependent on international research collaboration (Confraria & Godinho, 2015) and, as we have seen in Fig. 3, it is also the region which is most dependent on funding from non-African partners and philanthropic institutions. Therefore it is interesting to notice that it is the region where the disease burden and health research specialization show a greater alignment.

In Table 2 we can also observe that the disease burden specialization in Southern Africa and West & Central Africa are also positively and significantly associated with their research specialization in models 1, 2 & 3. However, when we include the lagged dependent variable (L_NSI_Pub) in the model (see model 5), the significance disappears. This means that the association between these two dimensions may be derived mostly from the existence of previous scientific capabilities in those areas and not so much from the awareness to the disease burden in their region.

In this regard, it is important to note that there is a high correlation between normalized relative scientific specialization NSI in 2011-2015 and previous scientific specialization in 2006-2010 in all regions (around 98%). Scientific activities are dominated by strong path-dependencies. If one country has scientists that are involved in a certain type of research it is very likely that they will continue to do their research in that area.

Finally, we find that all African regions also seem to be specialized in areas where they have higher levels of international collaboration. Since the research in most African countries is highly dependent on international research collaboration and international research funding, this was an expected result.

To assess the extent to which a higher share of funding from international donors in specific scientific areas/diseases is associated with a higher disease burden specialization, we compute an additional set of regressions that have as dependent variables the normalized relative specialization index of a certain funder type in a specific disease. As stated in Data and Methods section, this study covers five types of funding organizations, i.e. Public African, public non-African, philanthropic, corporation and non-funded (or non-identified).

Comparisons of results from the five sub-groups in Table 3 enable us to observe a set of findings: First, the disease fields in which public non-African and philanthropic organizations fund relatively more are the ones, on average, that have higher disease burden specialization, in Eastern Africa, Southern Africa and Western & Central Africa. As we have seen in Fig. 3, these three regions are highly dependent on international research funding. Therefore, this finding is significant since it shows that these international donors, on average, are funding research on diseases that are relevant to these African regions. In Northern Africa, there is a

negative association between disease burden specialization and research funding specialization in all funding categories.

Table 2. Regression analysis: Match between disease burden specialization and research specialization.

	NSRI_Pubs_11.15				
Ind. Variables	model (1)	model(2)	model(3)	model(4)	model(5)
NSI_DALYs_2010	0.76***				
	(0.10)				
NSI_DALYs_2010*Eastern_Africa		1.21***	1.20***	1.08***	0.11**
		(0.14)	(0.14)	(0.15)	(0.046)
NSI_DALYs_2010*Northern_Africa		-0.016	-0.047	-0.37	-0.042
		(0.29)	(0.28)	(0.29)	(0.053)
NSI_DALYs_2010*Southern_Africa		0.46*	0.53**	0.56***	0.065*
		(0.25)	(0.27)	(0.21)	(0.039)
NSI_DALYs_2010*West&Central_Africa		0.90***	0.91***	0.71***	0.048
		(0.11)	(0.11)	(0.14)	(0.055)
Int_collab_11_15 (%)				1.14***	0.16**
				(0.30)	(0.081)
L_NSI_Pubs					0.89***
					(0.022)
Constant	0.076**	0.058*	0.029	-0.90***	-0.078
	(0.031)	(0.029)	(0.062)	(0.26)	(0.064)
Regional dummies	No	No	Yes	Yes	Yes
Observations	103	103	103	103	103
R-squared	0.393	0.510	0.516	0.571	0.967

Note: 1) Robust standard errors in parentheses; *** p<0.01, ** p<0.05, * p<0.1. 2) In models 3-5, the regression model was computed controlling for four regions: Eastern Africa, Northern African, Southern Africa and West & Central Africa.

Table 3. Regression analysis: Match between disease burden specialization of a region and medical research specialization of a specific funder group.

	Research funded by African public organizations		**Research funded by non-African public organizations**		**Research funded by philanthropic organizations**		**Research funded by corporation**		**Un-funded Research**	
Ind. Variables	model(1)	model(2)	model(3)	model (4)	model(5)	model (6)	model(7)	model (8)	model(9)	model(10)
NSI_DALYs_2010*	1.06***	0.57**	1.26***	0.33***	1.40***	0.70***	1.16***	0.79***	0.86***	-0.056
Eastern_Africa	(0.16)	(0.22)	(0.17)	(0.089)	(0.13)	(0.15)	(0.15)	(0.15)	(0.16)	(0.081)
NSI_DALYs_2010*	-0.44	-0.39	-0.50*	-0.36***	-0.65**	-0.68**	-0.084	-0.49	0.17	0.071
Northern_Africa	(0.41)	(0.27)	(0.29)	(0.12)	(0.32)	(0.28)	(0.29)	(0.30)	(0.26)	(0.11)
NSI_DALYs_2010*	0.51*	0.29*	0.51	0.19**	0.66**	0.47***	0.60**	0.42***	0.58**	0.055
Southern_Africa	(0.29)	(0.15)	(0.35)	(0.073)	(0.32)	(0.17)	(0.24)	(0.095)	(0.26)	(0.055)
NSI_DALYs_2010*	0.74***	0.26	1.14***	0.17*	1.25***	0.56***	0.86***	0.35**	0.73***	-0.021
West&Central_Africa	(0.18)	(0.18)	(0.11)	(0.097)	(0.11)	(0.13)	(0.17)	(0.17)	(0.13)	(0.084)
Int_collab_11_15 (%)		0.32		0.44**		0.90***		0.91*		0.24*
		(0.37)		(0.18)		(0.32)		(0.48)		(0.14)
African_Funding		0.47***								
L_NSI_FundCat		(0.091)								
Non-African_Funding				0.72***						
L_NSI_FundCat				(0.053)						
Philanthropic_Funding						0.41***				
L_NSI_FundCat						(0.083)				
Corporation_Funding								0.38***		
L_NSI_FundCat								(0.066)		
Non-Funded										0.86***
L_NSI_FundCat										(0.049)
Constant	-0.015	-0.18	-0.035	-0.34**	-0.026	-0.70***	0.029	-0.58	-0.41***	-0.36***
	(0.077)	(0.31)	(0.072)	(0.15)	(0.067)	(0.27)	(0.076)	(0.40)	(0.060)	(0.12)
Regional effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	101	101	101	101	101	101	101	101	101	101
R-squared	0.367	0.584	0.503	0.899	0.544	0.728	0.375	0.624	0.651	0.934

Note: 1) Robust standard errors in parentheses; *** p<0.01, ** p<0.05, * p<0.1. 2) The regression model was computed controlling for four regions: Eastern Africa, Northern African, Southern Africa and West & Central Africa.

Second, Eastern Africa is the only region where a positive association exists, on average, between disease burden specialization and public African funding specialization. Third, unexpectedly publications funded by corporations seem to be positively associated with disease burden in all regions except for Northern Africa. Fourth, there is no clear association between research that is not funded by a specific institution and disease burden specialization (see model 10 in Table 3). The only driver of research that has no funding acknowledgements seems to be previous specialization on that topic (lagged dependent variable).

Finally, we should note that we controlled for the level of international collaboration and previous share of research funding (2009-2010) for all the dependent variables. As expected, we only find a significant positive association between the intensity of international collaboration and medical research funded by public non-African and philanthropic organizations (see models 4 & 6 in Table 3). This may happen because internationally collaborated research is usually supported by external research funding. As for the previous funding specialization in a specific disease, there is a positive association between our lagged variable and all dependent variables. This indicates the path dependence that exists in research funding in general. Investing requires confidence on the part of the investor that they will see a return on their investment. In environments where the logistics for research might be complex and challenging, the inclination is to fund governments and institutions with a track record of success (Head et al., 2017).

Discussion and Conclusions

While the vast majority of burden of disease globally is based in low and middle-income countries, only a small proportion of health research is performed in those regions. Therefore a common rationale is that these countries should use their limited resources to study diseases that are relevant to their health needs.

In this article, we evaluate the alignment between the medical research efforts and the burden of disease across four African regions. Within each region, we estimated the research and disease burden specialization (compared to the world specialization levels) across 28 diseases. Surprisingly, what we find is that in sub-Saharan Africa most diseases with a high disease burden are also the ones with relatively more research efforts. We find that the region with higher positive association between disease burden and research efforts is Eastern Africa. Northern Africa is the region where these two dimensions are less aligned. Contrary to what some literature suggests, our results indicate that the regions with higher dependence on public non-African and philanthropic funding institutions are the ones where the alignment between disease burden and research specialization is higher.

These findings are interesting for two main reasons. One, it has been argued that there are substantial misalignments, at the global level and at local levels, between research efforts and WHO estimates of health burden for a given disease (e.g. Evans et al., 2014; Rafols & Yegros, 2017). What we find is that while this may be true at the global level (high-income countries perform most of their medical research on diseases that are not the ones with a higher global disease burden), Sub-Saharan African researchers are performing research that is relevant for their regional health needs. Second, some authors also argue that researchers from high-income countries secure most of the funding for global health research projects in low-income regions, and often dictate the research agenda which leads to inappropriate projects unrelated to local research needs, and derive conclusions that do not have any direct local benefit (Binka, 2005; Gaillard, 1994). The results from this article seem to contradict this idea. What we find is that most international research funders (public non-African and philanthropic) support research on HIV/AIDS, tuberculosis and parasitic and vector diseases, which are diseases that generate a big share of disease burden in sub-Saharan Africa. What

our regression results seem to show is that, on average, high levels of dependence on international donors are not necessarily associated with less alignment between local health needs and local medical research efforts.

This does not mean that because international health research funders' priorities seem to be aligned with local African health needs, African countries do not need stronger scientific and institutional capacity. It is well known that substantial advantages exist in investment in local research, particularly with regard to ownership of the results, trust, inter-sector sharing of expertise between researchers and policy makers, and increased contextualisation of findings. Local knowledge from the people most directly affected by an issue is crucial to finding an innovative way of addressing it. Therefore, creating schools of public health and other institutions to train quality scientists in public health should continue to be a priority, as many African countries have few or no institutions that can provide proper training in public health research. Collaborative solutions and establishing funding partnerships between countries can also be a possibility for countries with fewer resources.

Our study has limitations and the results must be interpreted with caution since publications in WoS (or DALYs) are imperfect estimates of research efforts (health needs) in a specific disease and country. First, measurement of priorities in medical research with scientific publications associated to certain diseases is not straightforward because there is some health research related to health education approaches, beliefs related to health and prevention, quality and financing of healthcare, that is important for health outcomes and do not necessarily derive from research on certain diseases. Second, since scientific production and disease burden change over time, future studies should conduct a dynamic analysis of DALYs and publications to understand how the two dimensions evolve together. Third, future studies should also analyse the extent to which the research that is funded is actually used to contribute to health action.

Data and Methods

Data Sources. To identify medical priorities, we used Disability-Adjusted Life Years (DALYs) from World Health Organization (WHO) to measure the burden of disease. One DALY can be thought of as one lost year of healthy life, and the measured disease burden is the gap between a population's health status and that of a normative reference population (World Health Organization, 2017). To present the funded (and unfunded) research efforts, we rely on scientific research publications indexed in WoS. From the 'funding organization' field in each article, we identify whether the research was financially supported, and the name of the funder(s). We use the whole counting method (if a publication has two different types of funding institutions in their acknowledgements we count one publication for each funding institution/disease).

Publication records were assigned to a specific disease field by searches in abstracts and titles. We built a set of keywords that are strongly associated to a specific disease (or group of diseases) based on the ICD-9 codes[2] and previous research (Cardoso et al., 2014; Chapman et al., 2017; MSF, 2016). After building our queries, two external peer reviewers[3] reviewed the keywords for each one of our 28 disease categories. In total, we have 59,486 documents that were associated to at least one specific disease (28) and one African region (Eastern Africa, Northern Africa, Southern Africa or West & Central Africa).

Metrics. With the hypothesis that health needs in earlier years should drive the research agenda in later years, we compare the number of articles published between 2011-2015 with

[2] http://icd9.chrisendres.com/

[3] One of the reviewers is a PhD student in international health and development, who worked for five years as a nurse in epidemic contexts in several African countries. The other is a nurse with 15 years of experience and a MSc in health economics.

the disease burden in 2010. First, we count DALYs in each disease per region/period and number of publications in each disease per region/period. Then, since different diseases have different propensities to affect people and be researched, we also compute specialization indexes (SI) to assess the specialization of each disease in a given region. We do this by computing the revealed comparative index (Balassa, 1965):

$$SI_{rd} = \frac{P_{rd}/\sum_d P_{rd}}{P_d/\sum_d P_d} \qquad (1)$$

where P is the number of publications in region ***r*** in disease ***d.*** This index can be interpreted as a "comparative advantage". If a region ***r*** has a higher specialization in disease ***d,*** it means that region ***r*** has more scientific research focused on disease ***d***. Likewise, based on DALYs data, we also calculate the revealed specialization index for disease types in each region.

The definition of the index implies that its value is necessarily null or positive but is not bound by an upper limit. For this reason, we standardize this measure as follows:

$$NSI = \frac{(SI - 1)}{(SI + 1)} \qquad (2)$$

The threshold value of the normalized specialization index (NSI) remains zero, but the range is now [-1, +1]. The standardization is implemented for both publication and disease SI values.

Econometric approach. In this study, our primary research question is to understand whether disease burden specialization is associated with medical research specialization between different African regions across different diseases. To address this, in our multivariate regression analysis (OLS), we use scientific specialization (NSI Pubs) as our dependent variable, and disease burden specialization as our main independent variable. Since most African countries are highly dependent on international research collaboration, in our model we control for level of international collaboration. We also control for previous scientific specialization due to the path dependent nature of scientific production.

In Equation (3), *NSI_Pub* is the scientific specialization index in a certain region *r*, disease field *d* and period *t* (2011-2015). *NSI_DALY* is the disease burden specialization index in period t-1 (2010). *IC* is the percentage of internationally co-authored publications, *L_NSI_Pub* is a lagged dependent variable from the previous period (2006-2010), and *R* is a control for each of the four African regions. Finally, α is the constant, and ε is the unobserved residual.

$$NSI\_Pub_{rd} = \alpha + \mu NSI\_DALY_{rd} + \varphi IC_{rd} + L\_NSI\_Pub_{rd} + R_r + \varepsilon_{rd} \quad (3)$$

Since we are also interested in understanding if international funders are supporting medical research that is relevant for the health needs of African regions, we also compute a set of regressions that estimate what the relation between disease burden specialization and research funding specialization by donor category is. We conduct this analysis by using five different types of donor categories (dependent variables): 1) African public funding; 2) Non-African public funding (including multilateral funding); 3) Philanthropic funding; 4) Corporation funding; and 5) Non-funded research (or not identified).

$$NSI\_FundCat_{rd} = \alpha + \mu NSI\_DALY_{rd} + \varphi IC_{rd} + L\_NSI\_FundCat_{rd} + R_r + \varepsilon_{rd} \quad (4)$$

In Equation (4), *NSI_FundCat* is our specialization index (for each of the five funding categories) in a certain region *r*, disease field *d* and period *t* (2011-2015), and *L_NSI_FundCat* is a lagged dependent variable from the previous period (2006-2010). The rest variables (*NSI_DALY, IC*, *R* and α) are the same as those in Equation (3).

Acknowledgements

We are particularly grateful to Robin Cowan, Bart Verspagen, Tommaso Ciarli, Ismael Ràfols, Joanna Chataway, Michael Hopkins and Daniele Rotolo for their useful comments and suggestions. We also thank Rodrigo Costas for providing valuable data, and Charlotte Oliveira and Marco Salvador for reviewing our keywords queries for each disease. This article also benefited from comments made by participants of a SPRU workshop (Brighton, 2018) and UNU-MERIT internal conference (Maastricht, 2018). Any remaining error is ours.

References

AOSTI (African Observatory for Science, T. and I. (2014). *Assessment of Scientific Production in the African Union Member States 2005-2010*.

Atal, I., Trinquart, L., Ravaud, P., & Porcher, R. (2018). A mapping of 115,000 randomized trials revealed a mismatch between research effort and health needs in non–high-income regions. *Journal of Clinical Epidemiology*, *98*, 123–132. http://doi.org/10.1016/j.jclinepi.2018.01.006

Balassa, B. (1965). Trade Liberalisation and "Revealed" Comparative Advantage. *The Manchester School*, *33*(2), 99–123. http://doi.org/10.1111/j.1467-9957.1965.tb00050.x

Binka, F. (2005). Editorial: North-South research collaborations: a move towards a true partnership? *Tropical Medicine and International Health*, *10*(3), 207–209. http://doi.org/10.1111/j.1365-3156.2004.01373.x

Cardoso, A. L., Breugelmans, G., Manville, C., Chataway, J., Cochrane, G., James, S., … Murali, N. (2014). Africa Mapping - Current state of health research on poverty-related and neglected infectious diseases in sub-Saharan Africa, 42.

Chapman, N., Doubell, A., Oversteegen, L., Chowdhary, V., Rugarabamu, G., Zanetti, R., Ming, O. Borri, J. (2017). Neglected disease research and development: reflecting on a decade of global investment, 121.

Confraria, H., & Godinho, M. M. (2015). The impact of African science: a bibliometric analysis. *Scientometrics*, *102*(2), 1241–1268. http://doi.org/10.1007/s11192-014-1463-8

Evans, J. A., Shim, J.-M., & Ioannidis, J. P. A. (2014). Attention to Local Health Burden and the Global Disparity of Health Research. *PLoS ONE*, *9*(4), e90147. http://doi.org/10.1371/journal.pone.0090147

Gaillard, J. F. (1994). North-South research partnership: Is collaboration possible between unequal partners? *Knowledge and Policy*, *7*(2), 31–63. http://doi.org/10.1007/BF02692761

Head, M. G., Goss, S., Gelister, Y., Alegana, V., Brown, R. J., Clarke, S. C., Fitchett, J., Atun, R. Scott, J., Newell, M-L., Padmadas, S., Tatem, A. J. (2017). Global funding trends for malaria research in sub-Saharan Africa: a systematic analysis. *The Lancet. Global Health*, *5*(8), e772–e781. http://doi.org/10.1016/S2214-109X(17)30245-0

Lee, B. Y., Bacon, K. M., Bottazzi, M. E., & Hotez, P. J. (2013). Global economic burden of Chagas disease: a computational simulation model. *The Lancet Infectious Diseases*, *13*(4), 342–348. http://doi.org/10.1016/S1473-3099(13)70002-1.Global

Mouton, J. (2008). Africa's Science Decline. *Harvard International Review*, *30*(3), 46–51. http://doi.org/Article

MSF. (2016). *Clinical guidelines - Diagnosis and treatment manual*. *Clinical guidelines Diagnosis and treatment manual*.

Ndounga Diakou, L. A., Ntoumi, F., Ravaud, P., & Boutron, I. (2017). Published randomized trials performed in Sub-Saharan Africa focus on high-burden diseases but are frequently funded and led by high-income countries. *Journal of Clinical Epidemiology*, *82*. http://doi.org/10.1016/j.jclinepi.2016.10.008

Rafols, I., & Yegros, A. (2017). Is Research Responding to Health Needs? *SSRN Electronic Journal*. http://doi.org/10.2139/ssrn.3106713

Røttingen, J. A., Regmi, S., Eide, M., Young, A. J., Viergever, R. F., Ardal, C., Guzman, J., Edwards, D., Matlin, S., Terry, R. F. (2013). Mapping of available health research and development data: what's there, what's missing, and what role is there for a global observatory? *Lancet*, *382*(9900), 1286–1307. http://doi.org/10.1016/S0140-6736(13)61046-6

Sawyerr, A. (2014). African Universities and the Challenge of Research Capacity Development. *Journal of Higher Education in Africa / Revue de l'enseignement Supérieur En Afrique*. CODESRIA. http://doi.org/10.2307/24486132

Tijssen, R. J. W. (2007). Africa's contribution to the worldwide research literature: New analytical perspectives, trends, and performance indicators. *Scientometrics*, *71*(2), 303–327. http://doi.org/10.1007/s11192-007-1658-3

UNESCO (Ed.). (2015). *UNESCO science report: towards 2030*. Paris: UNESCO Publishing.

World Health Organization. (2012). *Research and Development to Meet Health Needs in Developing Countries : Strengthening Global Financing and Coordination Report of the Consultative Expert Working Group on Research and Development : Financing and Coordination Briefing to Representatives of*.

World Health Organization. (2017). *WHO methods and data sources for global burden of disease estimates*.

Comparing Coverage of Scopus, WoS, and OBRSS List: A Case for Institutional and National Databases of Research Output?

Lai Ma[1] and Liam Cleere[2]

[1]*lai.ma@ucd.ie*
School of Information and Communication Studies, University College Dublin, Dublin 4, Ireland

[2]*liam.cleere@ucd.ie*
Office of the Vice President for Research and Innovation, University College Dublin, Dublin 4, Ireland

Abstract

University College Dublin (UCD) has implemented the Output-Based Research Support Scheme (OBRSS) since 2016. Adapted from the Norwegian model, the OBRSS aims to incentivise research and publications by disbursing research support funds based on research outputs including publications and PhD supervision. This paper examines the coverage of UCD publications in Scopus, Web of Science (WoS), and the OBRSS list. The analysis shows that the OBRSS list has a more comprehensive coverage in all disciplines and has a significant advantage in the coverage of SSH disciplines compared to Scopus and WoS. It is also evident that the development of the OBRSS list is more transparent about its indexing practices and procedures. Further, the list offers opportunities for academic and research community to co-construct the list and to justify the prestige of publications, which allows the inclusion of novel, non-commercial, and open access publications without track records of citations and journal impact factor. The OBRSS has also encourages regular updates of publication records by academic and research staff, which leads to mostly complete information about publication activities useful for bibliometric analysis, research management, and strategic planning.

Introduction

The Output-Based Research Support Scheme (OBRSS) has been implemented in University College Dublin (UCD) since 2016. Adapted from the Norwegian model, the OBRSS aims to incentivise research and publication by disbursing research support funds based on research outputs including publications and PhD supervision. The design of the OBRSS involved the construction of a ranked publication list and a points system, and its implementation is contingent on regular and reliable updates on the Current Research Information System by academic and research staff.

At the time of this writing, there is not a national initiative to collate complete data of academic and research publications in Ireland, nor is there a systematic effort to create a database such as CRIStin in Norway (Sivertsen, 2018) or the Publication Forum in Finland (Pölönen, 2018). Commercial databases and tools such as Elsevier's Scopus and SciVal and Thomas Reuter's Web of Science (WoS) are used as data sources for counting research outputs and depicting publication trends, despite the common knowledge that the coverage of these commercial databases is incomplete, especially for disciplines in the social sciences and the humanities (Mongeon & Paul-Hus, 2016; Sivertsen & Larsen, 2012). Zacharewicz, et. al (2018) has shown that over 70% of research funding in Ireland has been channelled to project funding, which largely benefits STEM research. While there is a strong emphasis on STEM-oriented projects, many disciplines in the social sciences and humanities in Ireland are ranked higher in subject rankings internationally. There is, however, a lack of data about publication patterns and other research outputs about all disciplines.

Based on a European survey, Sīle, et. al (2018) show that not all EU member states maintain national bibliographic databases and the completeness and uses of existing ones vary. Bibliographic data can be linked to data reporting, research evaluation, and research funding

allocation in some countries but not others. It has also been noted that the COST Action ENRESSH (European Network for Research Evaluation in the Social Sciences and Humanities) envisions a European database by integrating existing databases and information systems in Europe (p. 12). The VIRTA-ENRESSH proof of concept pilot (Puuska, et al, 2018) and the Nordic List (NSD, 2018) are pioneering projects in this area.

What are the benefits of institutional/national bibliographic databases? Sivertsen (2010) has discussed the need for complete data when designing performance indicator. A few case studies have also been conducted to understand the validity of bibliometric analyses in local information systems such as METIS in the Netherlands (van Leeuwen, van Wijk & Wouters, 2016). There is a general agreement that national/institutional databases or information systems are necessary for bibliometric analysis in the social sciences and the humanities (see, for example, Sivertsen & Larsen, 2012; Ossenblok, Engels & Sivertsen, 2012) because the coverage of commercial databases is less than satisfactory in some disciplines, not to mention the indexing practices of commercial providers can be driven by market interests.

The uses of national/institutional databases are relatively under-explored in bibliometrics. This paper examines the coverage of UCD publications in Scopus, Web of Science, and the OBRSS list, followed by a discussion of the potential benefits of institutional and national databases for research evaluation and open science, as well as related issues pertaining to fairness and transparency of research management and research policy.

Background: OBRSS

University College Dublin (UCD) recognises that faculty's commitment to excellent research helps build a strong research reputation. The university also recognises that many of the day-to-day costs of research activity are not covered by research grants. In recognition of this, UCD has developed the Output-Based Research Support Scheme (OBRSS) to disburse research support funds to faculty based on their research outputs, as captured through publications and PhD supervision.

Publications are divided into two categories: normal and prestigious. The OBRSS uses lists of publication channels – one for Publishers and one for Series (Journals, Book Series, and Conference Series) – as a reference for the categorisation of publications. The lists are dynamic and are updated once every year by using the Danish, Finnish & Norwegian peer reviewed publication channel lists as a baseline. Refinements are made to this by UCD faculty each year.

The OBRSS uses the ranked publication list – one section for Publishers and another for Series (Journals, Book Series, and Conference Series) – as a reference for the calculation of points. Each publication is assigned one of two levels: level 1 – Normal or level 2 – Prestigious. Weighted scores are then applied to each publication. Similar to the Norwegian model, points are allocated for different types of publication as summarised in Table 1:

Table 1 Points allocation per publication type

Publication types	Points Level 1 'normal'	Points Level 2 'prestigious'
Book	5	8
Journals Article	1	3
Book Chapter	1	3
Conference Publication	0.5	2
Edited Book	1	3
Other Publication	0.5	2
Published Report	1	3

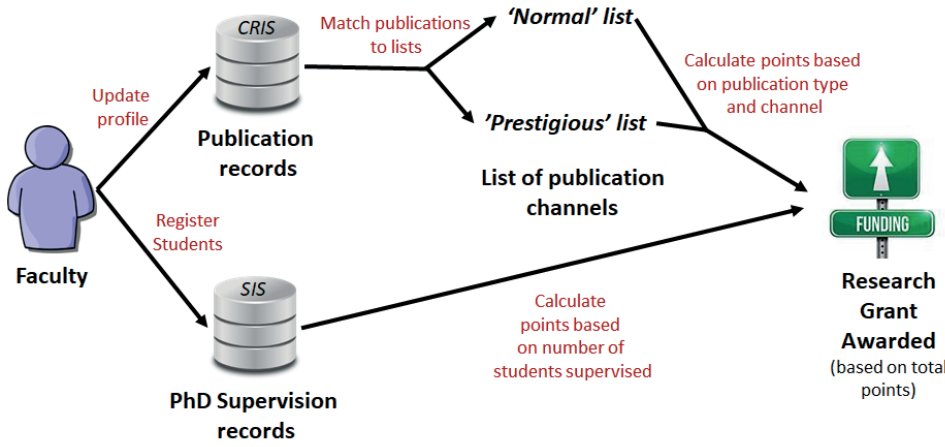


Figure 1 OBRSS Points

There is a consultation process to ensure that inputs from the academic staff are considered in finalising the ranked publication list. During the consultation period, academic staff can make recommendations to add/remove publications to/from the ranked publication list at the two levels. The suggestions and recommendations are reviewed by the Office of the Vice President for Research. Considering the objectives and scope of the OBRSS, external panels are not used to review the ranked publication list. Overall, approximately 1% of the total annual research budget for the university is allocated to the OBRSS.

Table 2 Calculation of publication output point

Publication output-points = B x C x F x N, where

- *B = Points (allocated based on the type of publications and whether it is in a 'normal' or 'prestigious' channel)*
- *C = collaboration factor (multiply by 1.25 if there are any international authors on the paper)*
- *F = UCD author factor (multiply by 0.7 if there are two UCD academic staff on the paper; multiply by 0.6 if there are three UCD academic staff on the paper; multiply by 0.5 if there are four or more UCD academic staff on the paper)*
- *N = if the total number of authors on a paper exceeds 100, multiply the result by 0.1*

Publication points are calculated for each academic staff's publications in the CRIS over a three-year period (for example 2015-2017) using the formula in Table 2. All academic staff are automatically entered into the OBRSS each year. The total points that an academic staff has accumulated is communicated using a personalised points statement. Final points statements are issued to academic staff receiving an award in October each year. The minimum value threshold for a research award is €200 and there is no maximum research award.

Coverage of Scopus, Web of Science and OBRSS List of UCD Research Outputs

One of the most significant outcomes of the implementation of the OBRSS is a more complete picture of publication records in University College Dublin. The number of academic staff updating their research profiles in the CRIS has increased each year. In the first year the OBRSS was implemented, 85% of academic staff updated their profiles as opposed to 75% over the previous three years. The publication records are essential to understand publication practices, in terms of publication types and frequency, for example, in different disciplines. While using the Danish, Finnish, and Norwegian lists as the baseline, the OBRSS list has been updated regularly with new publications recorded in CRIS, as well as inputs from academic staff.

The OBRSS ranked publications list covers over 78.6% of all UCD publications, with the highest at 88.7% in the College of Science and the lowest at 67.5% in the College of Business (Table 3). The OBRSS list is primarily designed to traditional publication outputs such as books and journal publications. Its coverage of journal articles is 97.9% while coverage of book-based publications average 80%. Since STEM disciplines publish more of their content on average in journals, this tends to improve the overall coverage for these disciplines.

Publications not 'counted' by the OBRSS list include 45% of conference papers; 80% of 'other publications' and 75% of published reports (Table 4). The lack of coverage is due to not having recognised publishers or unique identifiers (e.g. ISSNs) associated with the publications.

While the coverage of Scopus and WoS of UCD publications varies, with biggest differences in the College of Engineering and Architecture and a small difference in the College of Arts and Humanities, the rankings of coverage is the same, from highest to lowest coverage: Science, Health and Agricultural Sciences, Engineering and Architecture, Business, Social Sciences and Law, and Arts and Humanities.

The rankings by School (Table 5) give a more nuanced comparison of the coverage of Scopus and WoS. Publications in Computer Science, for example, are covered much higher in Scopus than WoS. The top ten subjects in both Scopus and WoS databases are STEM-related, with the exception of Economics in Scopus. It is clear that the coverage of arts and humanities subjects is very weak, as has been surveyed by many (see Mongeon & Paul-Hus, 2016 for a recent analysis). This is in stark contrast to the OBRSS coverage where there is much higher coverage in all disciplines including arts and humanities.

In summary, the analysis shows that there has been a lack of data sources for understanding publication trends in the social sciences and the humanities, at least in the context of UCD. The coverage and discrepancies between the two major commercial databases are not negligible. It is also clear that research evaluation at individual and institutional level (e.g., university rankings) should not depend solely on these databases, particularly taking into account the levels of coverage in each subject, not to mention the differences in citation practices that would affect h-index, for example. Overall, the OBRSS list represents a higher percentage of publications in all subjects, but has a significant advantage in the social sciences and the arts and humanities.

Table 3 Comparison of publication volumes per college for academic staff only from 2015 to 2017 inclusive; using CRIS/OBRSS data from UCD RMS Profiles June 2018, Scopus data from June 2018; and Web of Science data (WoS) from Jan 2019

UCD College Name	CRIS Total 2015-17	OBRSS Total 2015-17	Scopus Total 2015-17	WoS Total 2015-17	% Coverage in OBRSS	% Coverage in Scopus	% Coverage in WoS
Science	3,221	2,856	2,365	1,990	88.7%	73.4%	61.8%
Health and Agricultural Sciences	4,294	3,517	2,489	2,351	81.9%	58.0%	54.8%
Engineering and Architecture	2,469	1,834	1,545	1,048	74.3%	62.6%	42.4%
Arts and Humanities	893	634	134	147	71.0%	15.0%	16.5%
Social Sciences and Law	2,168	1,489	750	548	68.7%	34.6%	25.3%
Business	652	440	278	210	67.5%	42.6%	32.2%
Grand Total	**13,697**	**10,770**	**7,561**	**6,294**	**78.6%**	**55.2%**	**46.0%**

Table 4 Comparison of publication volumes per types for academic staff only from 2015 to 2017 inclusive; using CRIS/OBRSS data from UCD RMS Profiles June 2018, Scopus data from June 2018; and Web of Science data (WoS) from Jan 2019

UCD School Name	CRIS Total 2015-17	OBRSS Total 2015-17	Scopus Total 2015-17	WoS Total 2015-17	% Coverage in OBRSS	% Coverage in Scopus	% Coverage in WoS
Book	207	140	53	9	67.6%	25.6%	4.3%
Chapter	1,273	1,027	292	135	80.7%	22.9%	10.6%
Conference Paper	3,506	1,931	891	659	55.1%	25.4%	18.8%
Edited Book	209	184	28	5	88.0%	13.4%	2.4%
Journal article	7,541	7,308	6,296	5,469	96.9%	83.5%	72.5%
Other	725	144	1	17	19.9%	0.1%	2.3%
Report	236	36			**15.3%**	**0.0%**	**0.0%**
Grand Total	**13697**	**10770**	**7561**	**6294**	**78.6%**	**55.2%**	**46.0%**

Discussion and Conclusion

The OBRSS has the objective of incentivising publications in high-quality, international publication outlets; at the same time, it also encourages regular updates of publication records by academic and research staff, which leads to close to complete information about publication records within the university. Before the implementation of the OBRSS, data were incomplete as academic staff were less motivated to keep their research profiles up-to-date.

Table 5 Comparison of publication volume per school for academic staff 2015-2017 inclusive: using CRIS/OBRSS data from UCD RMS Profiles June 2018, Scopus data from June 2018; and Web of Science data (WoS) from Jan 2019

UCD School Name	CRIS Total 2015-17	OBRSS Total 2015-17	Scopus Total 2015-17	WoS Total 2015-17	% Coverage in OBRSS	% Coverage in Scopus	% Coverage in WoS
Physics	883	854	810	733	96.7%	91.7%	83.0%
Chemistry	272	254	229	236	93.4%	84.2%	86.8%
Biomolecular and Biomedical Science	434	396	309	295	91.2%	71.2%	68.0%
Medicine	1,621	1,471	1,066	926	90.7%	65.8%	57.1%
Biology and Environmental Science	304	275	205	183	90.5%	67.4%	60.2%
Economics	114	102	80	66	89.5%	70.2%	57.9%
Electrical and Electronic Engineering	574	507	489	300	88.3%	85.2%	52.3%
Music	44	38	6	5	86.4%	13.6%	11.4%
Philosophy	224	190	54	37	84.8%	24.1%	16.5%
Earth Sciences	283	239	84	91	84.5%	29.7%	32.2%
Chemical and Bioprocess Engineering	163	134	130	107	82.2%	79.8%	65.6%
Veterinary Medicine	736	603	406	432	81.9%	55.2%	58.7%
Mathematics and Statistics	426	349	291	204	81.9%	68.3%	47.9%
Biosystems and Food Engineering	415	334	278	228	80.5%	67.0%	54.9%
Computer Science	619	489	437	248	79.0%	70.6%	40.1%
Public Health, Physiotherapy and Sports Science	676	534	354	380	79.0%	52.4%	56.2%
English, Drama and Film	252	192	55	59	76.2%	21.8%	23.4%
Sociology	138	104	48	45	75.4%	34.8%	32.6%
Politics and International Relations	182	137	77	63	75.3%	42.3%	34.6%
Languages, Cultures and Linguistics	210	157	37	36	74.8%	17.6%	17.1%
Archaeology	118	88	23	13	74.6%	19.5%	11.0%
Agriculture and Food Science	945	688	537	504	72.8%	56.8%	53.3%
Mechanical and Materials Engineering	609	440	348	195	72.2%	57.1%	32.0%
History	200	142	28	39	71.0%	14.0%	19.5%
Nursing, Midwifery and Health Systems	316	221	126	109	69.9%	39.9%	34.5%
Psychology	389	267	176	157	68.6%	45.2%	40.4%
Civil Engineering	351	238	194	132	67.8%	55.3%	37.6%
Art History and Cultural Policy	77	52	3	7	67.5%	3.9%	9.1%
Business	652	440	278	210	67.5%	42.6%	32.2%
Education	100	67	37	23	67.0%	37.0%	23.0%

Classics	24	16	4		66.7%	16.7%	0.0%
Information and Communication Studies	98	64	48	20	65.3%	49.0%	20.4%
Law	306	182	35	12	59.5%	11.4%	3.9%
Geography	230	133	95	80	57.8%	41.3%	34.8%
Social Policy, Social Work and Social Justice	269	155	77	32	57.6%	28.6%	11.9%
Architecture, Planning and Environmental Policy	357	181	106	86	50.7%	29.7%	24.1%
Irish, Celtic Studies and Folklore	86	37	1	1	43.0%	1.2%	1.2%
Grand Total	**13,697**	**10,770**	**7,561**	**6,294**	**78.6%**	**55.2%**	**46.0%**

The OBRSS ranked publication list also gives a better overview of research activities of all disciplines, compared to data generated from commercial providers such as Scopus or WoS. This study shows that the coverage of the OBRSS list is significantly more comprehensive in SSH disciplines. These data can lead to better strategic planning, research management and research policy. For example, UCD School of History is ranked in the top 100 in the QS 2019 subject rankings, but only 14% of its publications were indexed by Scopus and 19.5% indexed by WoS. The data generated by the institutional database would be more useful to analyse the publication trends and other factors that contribute to high academic reputation, for example.

National and institutional databases with comprehensive metadata can also allow for analyses based on factors such as gender and career stage. Currently, most commercial databases and services such as ORCID do not record gender, career stage, and other useful factors for bibliometric analyses. The lack of these data makes analyses complicated and prone to errors.

While an ideal list of publications is difficult to attain, the OBRSS list, like the Norwegian and Finnish list, offers opportunities for academics and researchers to make suggestions and comments. Hence, the prestige of a journal does not solely depend on the journal impact factor, or it being indexed by commercial databases; rather, to a certain extent, the prestige of publications is a consensus of the academic and research community. Ideally, disagreement and discontentment are resolved in open discussion with the goal of maintaining a fair and representative list of publications.

The co-construction of an institutional and national list also allows the inclusion of new publications that have not accumulated citations and hence have not been indexed by commercial databases. These could include publishers who are highly recommended by experts, and those who are enthusiastic about open science and open access. The inclusion of new publishers, apart from the Big Deals, would be beneficial to knowledge production as it allows new voices to be heard—and be rewarded.

Currently, research profiles involving individual and institutional research performances are usually generated by commercial services such as Scopus and Google Scholar. However, their indexing practices are not transparent compared to national or institutional databases, for example, in the case of the Norwegian list (Sivertsen, 2010) and, for instance, the OBRSS list. National and institutional databases could be a trusted source with fair and transparent procedures with inputs from the academic and research community.

To conclude, the construction of the OBRSS list would not be possible without sufficient resources and support. The implementation of the OBRSS encourages regular updates on publication records, which also assists refinement of the ranked publications list. However, like other performance-based funding systems, the effects on publication trends and research practices would require examination over time (see, for example, Aagaard, 2015; Butler, 2003; Hammerfelt & de Rijcke, 2015; Hicks, 2012, Ma, 2018). Nevertheless, the comparison of coverage shows that there are benefits of institutional and national databases. First, the publications represent institutional and national areas of interest and there is better coverage of publications in SSH disciplines. Second, the databases can provide quantitative evidence to support qualitative peer-review assessments, particularly in research areas not sufficiently indexed in Scopus or WoS. Third, the databases can be co-constructed by the academic and research community and are hence more fair and transparent for research evaluation and other uses. Last, national and institutional databases can also be more open to recommendations of new, open access publications. This study provides some supporting evidence for the vision of a European databases (see Puuska, et al., 2018) by comparing the coverage of the Scopus, WoS, and the OBRSS list.

Acknowledgements

Dr. Lai Ma would like to thank Professor Gunnar Sivertsen for his generous support during her STSM visit to Nordic Institute for Studies in Innovation, Research and Education (NIFU) in Oslo, Norway. This work has been conducted within the framework of the COST action "European Network for Research Evaluation in the Social Sciences and Humanities" (ENRESSH, CA15137, enressh.eu).

References

Aagaard, K. (2015). How incentives trickle down: Local use of a national bibliometric indicator system. *Science and Public Policy*, *42*, 725–737.

Butler, L. (2003). Modifying publication practices in response to funding formulas. *Research Evaluation*, 12(1), 39–46. http://doi.org/10.3152/147154403781776780

Engels, T. C. E., & Guns, R. (2018). The Flemish performance-based research funding system: A unique variant of the Norwegian model. *Journal of Data and Information Science*, *3*(4), 44-59. DOI: 10.2478/jdis-2018-0020

Hammarfelt, B., & de Rijcke, S. (2015). Accountability in context: Effects of research evaluation systems on publication practices, disciplinary norms, and individual working routines in the faculty of Arts at Uppsala University. Research Evaluation, 24(1), 63–77.

Hicks, D. (2012). Performance-based university research funding systems. Research Policy, 41(2), 251–261. http://doi.org/10.1016/J.RESPOL.2011.09.007

Ma, L. Responses to Output-Based Research Support Scheme in University College Dublin, 23rd International Conference on Science and Technology Indicators (STI 2018), Leiden, The Netherlands, September 12-14, 2018.

Mongeon, P., & Paul-Hus, A. (2016). The journal coverage of Web of Science and Scopus: A comparative analysis. *Scientometrics*, *106*, 213-228. DOI: 10.1007/s11192-015-1765-5

NSD. (2018). The Nordic List. Retrieved from https://dbh.nsd.uib.no/publiseringskanaler/Forside.action?request_locale=en

Ossenblok, T. L. B., Engels, T. C., & Sivertsen, G. (2012). The representation of the social sciences and humanities in the Web of Science – a comparison of publication patterns and incentive structures in Flanders and Norway (2005–9). Research Evaluation, 21(4), 280–290.

Pölönen, J. (2018). Applications of, and experiences with, the Norwegian Model in Finland. *Journal of Data and Information Science*, *3*(4), 30-43. DOI: 10.2478/jdis-2018-0019

Puuska, H-M., Engels, T., Guns, R., Sivertsen, G., Pölönen, J., & Jorge, M-R. (2018). Proof of concept of a European database for social sciences and humanities publications: The VIRTA-ENRESSH pilot. CRIS2018: 14th International Conference on Current Research Information Systems, Umeå, June 13-16, 2018.

Schneider, J. W., Aagaard, K., Block, C. (2016). What happens when national research funding is linked to differentiated publication counts? A comparison of the Australian and Norwegian publication-based funding models. *Research Evaluation*, *25*(3), 244-256. http://doi.org/10.1093/reseval/rvv036

Sīle, L., et al. (2018). Comprehensiveness of national bibliographic databases for social sciences and humanities: Findings from a European survey. Research Evaluation, forthcoming. Retrieved from https://doi.org/10.1093/reseval/rvy016

Sivertsen, G. (2010). A performance indicator based on complete data for the scientific publication output at research institutions. ISSI Newsletter, 6(1), 22–28.

Sivertsen, G. (2018). The Norwegian Model in Norway. *Journal of Data and Information Science*, *3*(4), 2-18. DOI: 10.2478/jdis-2018-0017

Sivertsen, G. & Larsen, B. (2012). Comprehensive bibliographic coverage of the social sciences and humanities in a citation index: An empirical analysis of the potential. Scientometrics, 91(2), 567–575.

van Leeuwen, T.N., van Wijk, E., Wouters, P.F. (2016). Bibliometric analysis of output and impact based on CRIS data: a case study on the registered output of a Dutch university. Scientometrics, 106(1), 1–16.

Zacharewicz, T., Lepori, B., Reale, E., & Jonkers, K. (2018). Performance-based research funding in EU Member States—a comparative assessment. Science and Public Policy. http://doi.org/10.1093/scipol/scy041

The regional balance of knowledge flows

Giovanni Abramo[1] and Ciriaco Andrea D'Angelo[2]

[1] *giovanni.abramo@uniroma2.it*
Laboratory for Studies in Research Evaluation, Institute for System Analysis and Computer Science (IASI-CNR), National Research Council of Italy (Italy)

[2] *dangelo@dii.uniroma2.it*
Department of Engineering and Management, University of Rome "Tor Vergata" (Italy)

Abstract
This work applies a new approach to measure knowledge spillovers. Assuming that citation linkages between articles imply a flow of knowledge from the cited to the citing authors, we investigate the geographic flows of scientific knowledge among Italian regions at both overall and field level. Compared to other contributions proposed in the literature, the recourse to publication citations rather than patent citations offers more precise and robust results. Findings show that larger regions in terms of research output are more likely net exporters of new knowledge. At the same time, we register a positive correlation between the share of intraregional spillovers and the size of overall scientific output of a region.

Introduction

The 2018 Nobel Prize in economics was awarded to Paul Romer for his contributions to the theory of long-run economic growth with "endogenous" technological change (Romer, 1986; 1990). Romer's endogenous growth theory ties the development of new ideas to the number of people working in the knowledge sector (i.e. an effort devoted to R&D). These new ideas make everyone else producing regular goods and services more productive – that is, ideas increase total factors productivity. The reason for that is a peculiar feature of ideas, the fact that they are "non-rival" (meaning that one's use of an idea, like a recipe or a mathematical formula, does not prevent somebody's else use of it). In theory, public knowledge (as that encoded in publications) can be shared endlessly. However, in practice the diffusion of knowledge decays with the geographical distance from the idea generator. Since knowledge transfer cannot be observed directly (Jaffe, Trajtenberg, & Fogarty, 2000), one relies on proxy measures, notably citations. Jaffe, Trajtenberg, and Henderson (1993) compared the geographic location of patent citations with that of the cited patents, to investigate the extent to which knowledge spillovers are geographically localized. They found that citations to domestic patents are more likely to be domestic, and more likely to come from the same state as the cited patents.
Assuming that citation linkages between articles imply a flow of knowledge from the cited to the citing authors (Mehta, Rysman, & Simcoe, 2010; Van Leeuwen & Tijssen, 2000), some scholars relied on publication citations to investigate the international geographic flows of scientific knowledge. Rabkin, Eisemon, Lafitte-Houssat, and McLean Rathgeber (1979) explored world visibility for four departments (botany, zoology, mathematics, and physics) of the universities of Nairobi (Kenya) and Ibadan (Nigeria). At the level of the single field, Stegmann and Grohmann (2001) measured knowledge "export" in the Dermatology & Venereal Diseases category of the 1996 CD-ROM Journal Citation Reports (JCR), and in seven dermatology journals not listed in the 1996 JCR. Hassan and Haddawy (2013) mapped knowledge flows from the United States to other countries in the field of Energy over the years 1996-2009. Abramo and D'Angelo (2018) trailed international spillovers of knowledge produced in Italy, in over 200 fields, by analysing publication citations. Abramo, D'Angelo, and Carloni (2018) conceptualized the "balance of knowledge flows" (BKF) at the international level. Among others, the authors measured the share of domestic vs foreign flows generated by a country's research system, by field and as compared to other countries.

To the best of our knowledge, there are no works trailing the domestic spillovers of knowledge by publication citations. This study intends to close the gap. The recourse to publication citations rather than patent citations offer more precise and robust results, as the order of magnitude of the number of publications is much higher than that of patents.
We use a bibliometric approach, assuming that all new knowledge produced is measured by publications indexed in bibliographic repertories. We also assume that citations are proxies of scholarly impact, i.e. when a publication is cited it has had an impact on scientific advancement because other scholars have drawn on it, more or less heavily, for the further advancement of science. All limitations and assumptions typical of bibliometric analyses then apply. As for publications, it must me noted that not all new knowledge is encoded in publications (e.g. tacit knowledge), and not all publications are indexed in bibliographic repertories. Furthermore, stating that citations certify knowledge spillovers does not imply that there are no exceptions, rather that it is the norm. Citations in fact are not always certification of real flows and representative of all flows. Uncitedness, undercitation, and overcitation may actually occur. Finally, citation-based analysis is unable to capture flows outside the scientific system, such as that of practitioners (e.g. a physician applying a new pharmacological protocol after reading relevant literature), students, or industry.[i]
This work intends to measure the spillover of knowledge across regions. Identifying the administrative regions of all world institutions publishing in a period of time is a formidable task. For that reason, we restrict our analysis to the national level, and in particular to the authors' country, Italy. While results cannot be generalized to other countries, the study can be easily replicated in other national contexts, and provide useful information to the policy maker, such as the share of intra- vs extra-regional knowledge spillovers generated by a region's research system, by field and as compared to other regions. A very high intra-regional share is expected in those fields where research is context specific or mainly oriented towards intra-regional needs.
We observe Italian publications indexed in Web of Science (WoS) in 2010-2012, and their citing domestic publications up to 31/05/2017,[ii] to measure the outflows of knowledge produced in a region to other regions in Italy, and the inflows of knowledge produced by other regions in a region. The latter allows us to set up a region's balance of knowledge flows (RBKF), which would register a surplus when the difference between knowledge outflows and inflows is positive, a deficit when the opposite is true. We measure also the share of intra- vs extra-regional spillovers within each of the 20 administrative regions in Italy.[iii]
In the next section we present the data and method of analysis. Section 3 provides the results from the elaborations both at overall and at field level. Section 4 closes the work with our considerations on the relevance of the study.

Data and method

In the period 2010-2012, the seven main databases of the WoS *core collection*[iv] indexes 255399 publications showing Italian affiliations
To measure the regional outflows of knowledge one needs to identify the region of production of the cited publications. Because of increasing research collaboration at both national and international level, identifying the region of production of a publication is not straightforward. Various approaches for assigning an inter-regional authored publication to a region can be envisaged: i) to each region the institutions in the address list belong; ii) to one single region, based on the frequency the authors of that region (or the institutions of that region), occur in the address list; or based on the affiliation of the corresponding author, or first and last authors in non-alphabetically ordered bylines; iii) fractionalizing the publication by the number of regions, institutions or authors.

The convention we adopt here is the following: We define a publication as "made in" a region if at least 50% of its co-authors are affiliated to organizations located in that region. Because we are dealing with domestic knowledge flows, we had to exclude publications produced abroad, i.e. those with more than half co-authors affiliated to foreign institutions. Furthermore, 17401 publications are both multi-authored and lack the author-affiliation link. We were forced to exclude those as well from the analysis.
The final dataset of analysis is then composed of 167630 "made in" Italy publications (Table 1). Of them, 163395 are "made in" single regions only, and 4235 in two regions.[v] To identify and assign the region of production to publications, we have developed a matching application of geographic metadata ("affil_CITY", "affil_PROVINCE", and "affil_ZIP_CODE"), as reported in the address of publications. We have run the same application also to identify the region(s) of production of the citing publications.
To account for multiple affiliations of authors, we adopt a fractional counting method. In case of authors with *m* different affiliations, we assign 1/*m* to each of her or his bibliometric addresses. Of course, the issue reveals critical for authors affiliated to institutions located in different countries and/or regions. To exemplify, consider the publication with WoS code 000309458000001.


Nonlinear dynamics of beta-induced Alfven eigenmode driven by energetic particles

By: Wang, X (Wang, X.)[1]; Briguglio, S (Briguglio, S.)[2]; Chen, L (Chen, L.)[1,3]; Di Troia, C (Di Troia, C.)[2]; Fogaccia, G (Fogaccia, G.)[2]; Vlad, G (Vlad, G.)[2]; Zonca, F (Zonca, F.)[1,2]

Author Information
Reprint Address: Wang, X (reprint author)
+ Zhejiang Univ, Inst Fus Theory & Simulat, Hangzhou 310027, Zhejiang, Peoples R China.
Addresses:
+ [1] Zhejiang Univ, Inst Fus Theory & Simulat, Hangzhou 310027, Zhejiang, Peoples R China
+ [2] Assoc Euratom ENEA Fus, I-00044 Frascati, Italy
+ [3] Univ Calif Irvine, Dept Phys & Astron, Irvine, CA 92697 USA

Its seven co-authors are affiliated to three different institutions, located in China, Italy and the U.S. Chen L. and Zonca F. show double affiliation. The Italian institution ENEA-Euratom (localized in the region Latium), scores 4.5 authorships out of 7, because 4 authors (Briguglio S., Di Troia C., Fogaccia G., Vlad G.) are affiliated solely to it and one, Zonca F., is affiliated also to another institution. Therefore, the publication is defined as "made in" Latium. To measure inter-regional spillovers, we replicate to the regional level the approach detailed in Abramo & D'Angelo (2018). When a publication is cited it has given rise to a "benefit". The number of "benefits" deriving from a publication equals the number of citations, and if the citing publication is co-authored by scholars from one or more regions, the benefit has crossed a regional administrative boundary. In the case of a citing publication whose address list shows institutions located in *p* different regions, the same benefit (citation) is "gained" contemporaneously by *p* regions, so we can say that it has given rise to *p* equal "gains", one for each region of the Italian institutions listed in the affiliation list of the citing publication. A publication cited by *q* other publications would give rise to *q* benefits and *q* x *p* gains. Among the *p* citing regions there could be also the region the cited publication is made in. In this case we define the relevant gain as "intra-regional".
To exemplify, consider the publication with WoS code 000209048200010:

Chinski, A., Foltran, F., Gregori, D., Passali, D., & Bellussi, L. (2011). Foreign bodies in the ears in children: The experience of the buenos aires pediatric ORL clinic. *Turkish Journal of Pediatrics, 53*(4), 425-429.

Such publication is classified as "made in" Tuscany since three of its five co-authors (Foltran, F.; Passali, D.; Bellussi, L.) belong to two institutions (University of Siena and University of Pisa) located in that region. At 31/05/2017 the publication has accrued nine citations, 6 of which by publications with at least one Italian address as detailed below.

Foltran, F., Passali, F. M., Berchialla, P., Gregori, D., Pitkäranta, A., Slapak, I., . . . Raine, C. (2012). Toys in the upper aerodigestive tract: New evidence on their risk as emerging from the susy safe study. *International Journal of Pediatric Otorhinolaryngology, 76*(SUPPL. 1), S61-S66.
Italian regions associated to authors' affiliations: Friuli Venezia Giulia; Latium; Piedmont; Tuscany; Veneto

Gregori, D., et al. (2012). The susy safe project overview after the first four years of activity. *International Journal of Pediatric Otorhinolaryngology, 76*(SUPPL. 1), S3-S11.
Italian regions associated to authors' affiliations: Campania; Emilia Romagna; Friuli Venezia Giulia; Latium; Piedmont; Tuscany; Veneto

Moretti, C., & Foltran, F. (2012). Prevention and early recognition: The role of family pediatrician. *International Journal of Pediatric Otorhinolaryngology, 76*(SUPPL. 1), S39-S41.
Italian regions associated to authors' affiliations: Veneto

Sarafoleanu, C., Ballali, S., Gregori, D., Bellussi, L., & Passali, D. (2012). Retrospective study on romanian foreign bodies injuries in children. *International Journal of Pediatric Otorhinolaryngology, 76*(SUPPL. 1), S73-S75.
Italian regions associated to authors' affiliations: Friuli Venezia Giulia; Tuscany; Veneto

Sih, T., Bunnag, C., Ballali, S., Lauriello, M., & Bellussi, L. (2012). Nuts and seed: A natural yet dangerous foreign body. *International Journal of Pediatric Otorhinolaryngology, 76*(SUPPL. 1), S49-S52.
Italian regions associated to authors' affiliations: Abruzzo; Friuli Venezia Giulia; Tuscany

Slapak, I., Passali, F. M., Passali, G. C., Gulati, A., Gregori, D., Foltran, F., . . . Raine, C. (2012). Non food foreign body injuries. *International Journal of Pediatric Otorhinolaryngology, 76*(SUPPL. 1), S26-S32.
Italian regions associated to authors' affiliations: Latium

In brief, this publication generates six domestic benefits and 20 domestic gains, four of which intra-regional (Tuscany-Tuscany).
The RBKF is constructed for each region measuring the gains associated to the inflows and outflows of knowledge among the 20 Italian regions.
The overall publications in the 2010-2012 period, and the relevant benefits and gains per Italian region are shown in Table 1. The first row shows, for example, that researchers from Abruzzo in the three-year under observation authored 6541 publications, 2856 of which "made in", since at least 50% of their coauthors listed in the byline are affiliated to institutions located in that region. In turn, 72.1% of said publications are cited (at 31/05/2017) generating a total of 8552 domestic benefits, with an average of 4.15 domestic benefits per "made in" cited publication (8552/2060). On average 1.63 regions appropriate such benefits for a total of 13973 gains (8552*1.63), of which 39.6% intraregional, i.e. related to citing publications authored by other researchers from the Abruzzo region. The share of intraregional gains varies from a minimum of 24% in the smallest region, Valle D'Aosta, to a max of 54.4% in Sicily, one of the two island regions.
In general, there is a positive correlation (Spearman $\rho = 0.608$) between the share of intraregional gains and the size of overall scientific production of a region. This can be due to the fact that in (scientifically) large regions it is likely to find large research laboratories/groups conducting research on topics of common interest.

Table 1: 2010-2012 publications, citations, benefits and gains at 31/05/2017, by region

Region	Total publications	"Made in" publications	Of which cited (a)	Total domestic benefits (No. of citations from Italian regions) (b)	Average domestic benefits per cited publication (b/a)	Total domestic gains (c)	Of which intra-regional	Average domestic gains per benefit (c/b)
Abruzzo	6541	2856 (43.7%)	2060 (72.1%)	8552	4.15	13973	5529 (39.6%)	1.63
Basilicata	1858	758 (40.8%)	582 (76.8%)	2384	4.10	3637	1567 (43.1%)	1.53
Calabria	5581	2962 (53.1%)	2258 (76.2%)	10443	4.62	15165	7751 (51.1%)	1.45
Campania	21365	12632 (59.1%)	9134 (72.3%)	45305	4.96	64386	34349 (53.3%)	1.42
Emilia Romagna	31773	16491 (51.9%)	11780 (71.4%)	51856	4.40	77055	36692 (47.6%)	1.49
Friuli Venezia Giulia	11907	4830 (40.6%)	3549 (73.5%)	15812	4.46	23549	10500 (44.6%)	1.49
Latium	48519	26485 (54.6%)	18643 (70.4%)	79858	4.28	120646	58770 (48.7%)	1.51
Liguria	10808	5054 (46.8%)	3516 (69.6%)	14660	4.17	22325	10197 (45.7%)	1.52
Lombardy	56236	31311 (55.7%)	21543 (68.8%)	95641	4.44	140911	70500 (50.0%)	1.47
Marche	5740	2960 (51.6%)	2150 (72.6%)	9040	4.20	14037	6042 (43.0%)	1.55
Molise	1529	377 (24.7%)	288 (76.4%)	1286	4.47	2347	771 (32.9%)	1.83
Piedmont	20437	11015 (53.9%)	7731 (70.2%)	33636	4.35	47614	23559 (49.5%)	1.42
Puglia	13106	6923 (52.8%)	5060 (73.1%)	21859	4.32	32242	15386 (47.7%)	1.47
Sardinia	6009	2959 (49.2%)	2198 (74.3%)	9202	4.19	13369	6605 (49.4%)	1.45
Sicily	16341	9722 (59.5%)	7306 (75.1%)	34648	4.74	48414	26327 (54.4%)	1.40
Tuscany	31204	16894 (54.1%)	11767 (69.7%)	53209	4.52	78646	38210 (48.6%)	1.48
Trentino Alto Adige	6224	2969 (47.7%)	2178 (73.4%)	9313	4.28	12875	6704 (52.1%)	1.38
Umbria	5560	2477 (44.6%)	1913 (77.2%)	10233	5.35	15469	7217 (46.7%)	1.51
Valle D'Aosta	184	53 (28.8%)	36 (67.9%)	91	2.53	179	43 (24.0%)	1.97
Veneto	24158	12137 (50.2%)	8675 (71.5%)	37291	4.30	54164	26139 (48.3%)	1.45

Results and analysis

The regional balance of knowledge flows at overall level

The RBKF of all 20 Italian regions is shown Table 2. As an example, we presents the results concerning the Latium region. As shown in Table 1, the "made in" Latium cited publications are 18643. Such publications generate a total of 120646 gains, 58770 of which appropriated by Latium institutions. The remaining 61876 (as shown in column 2 of Table 2) are appropriated by the other 19 regions, which in turn publish altogether 84693 cited publications, generating a total of 352283 domestic gains, of which Latium appropriates 55862 (15.9%). The Latium RBKF is therefore positive (surplus) and equal to +6014, given the imbalance between the generated domestic gains (61876) and the earned gains (55862). The RBKF is positive also for Lombardy (+17386), Tuscany (+4738), Piedmont (+1710) and Campania (+802), i.e. the largest regions in terms of scientific production. The only exception among the largest region is Emilia Romagna showing a negative RBKF (-1621), as do the remaining 14 regions. In general, there is a positive correlation (Spearman $\rho = 0.849$) between the size of scientific production of a region and the value of its RBKF.

Table 3 shows complete data related to the flows at issue. Data on the main diagonal of the matrix illustrate the share of gains generated by a region which remain within that region (intraregional gains).[vi]

The matrix must be read by row, because the row vector shows the regional flows of knowledge produced in a given region (summing up to 100%). To exemplify, 16.2% of knowledge flows

out of Abruzzo are appropriated by Latium, the double as much as by Lombardy (8.0%), followed by Emilia Romagna (5.7%) and Tuscany (4.5%).
Needless to say, the matrix may be read by column also, in which case it will show an insight into the origin of the knowledge inflows of a given region (of course, in this case values should be rescaled, since the column does not sum up to 100%).

Table 2: The RBKF (among parenthesis percentages out of total gains)

Region	Extra-regional gains generated (a)	Cited extra-regional publications	Extra-regional gains generated by extra-regional publications	Earned gains (b)	RBKF (a-b)
Abruzzo	8444 (60.4%)	97992	394856	13289	-4845
Basilicata	2070 (56.9%)	99210	404337	3808	-1738
Calabria	7414 (48.9%)	97799	398397	9748	-2334
Campania	30037 (46.7%)	92021	378910	29235	+802
Emilia Romagna	40363 (52.4%)	90121	366161	41984	-1621
Friuli Venezia Giulia	13049 (55.4%)	96802	391831	16314	-3265
Latium	61876 (51.3%)	84693	352283	55862	+6014
Liguria	12128 (54.3%)	96891	393656	14489	-2361
Lombardy	70411 (50.0%)	82326	355120	53025	+17386
Marche	7995 (57.0%)	97951	398897	9248	-1253
Molise	1576 (67.1%)	99457	403568	4577	-3001
Piedmont	24055 (50.5%)	93468	385800	22345	+1710
Puglia	16856 (52.3%)	95506	389875	18270	-1414
Sardinia	6764 (50.6%)	97892	399111	9034	-2270
Sicily	22087 (45.6%)	93610	385597	22548	-461
Tuscany	40436 (51.4%)	90236	372447	35698	+4738
Trentino Alto Adige	6171 (47.9%)	97999	399015	9130	-2959
Umbria	8252 (53.3%)	98079	397977	10168	-1916
Valle D'Aosta	136 (76.0%)	99683	407630	515	-379
Veneto	28025 (51.7%)	92716	379287	28858	-833

Table 3: Overall import-export of knowledge (percentages out of total gains) between regions

		Earning																			
	Region	Abruzzo	Basilicata	Calabria	Campania	Emilia Romagna	Friuli Venezia Giulia	Latium	Liguria	Lombardy	Marche	Molise	Piedmont	Puglia	Sardinia	Sicily	Tuscany	Trentino Alto Adige	Umbria	Valle D'Aosta	Veneto
Generating	Abruzzo	39.6	0.3	0.7	4.1	5.7	1.7	16.2	1.3	8.0	1.6	0.7	2.2	3.1	0.9	3.0	4.5	0.4	2.8	0.0	3.4
	Basilicata	1.3	43.1	2.0	9.7	4.1	2.0	8.0	0.7	4.4	0.9	0.8	1.4	7.7	1.0	4.0	3.8	0.7	2.5	0.1	1.8
	Calabria	1.3	0.7	51.1	5.9	4.4	1.0	7.9	2.5	5.1	0.9	0.3	1.7	2.0	1.4	6.6	3.7	0.6	1.1	0.0	1.8
	Campania	1.6	1.1	1.6	53.3	4.1	1.3	8.5	1.5	5.9	1.1	1.0	2.1	2.6	0.9	3.3	5.6	0.6	1.4	0.0	2.3
	Emilia Romagna	1.5	0.3	0.8	3.3	47.6	2.6	8.5	1.8	9.6	1.4	0.3	3.1	2.2	1.1	3.1	5.1	1.1	1.1	0.0	5.4
	Friuli Venezia Giulia	1.5	0.3	0.9	3.0	6.5	44.6	8.0	1.5	8.6	1.5	0.2	3.0	2.3	0.9	2.7	5.2	1.9	0.9	0.1	6.7
	Latium	3.7	0.6	1.4	4.7	5.6	2.2	48.7	1.7	8.4	1.3	1.1	2.6	2.5	1.3	3.1	5.2	0.8	1.6	0.0	3.4
	Liguria	1.2	0.1	1.7	4.0	5.5	1.6	7.2	45.7	9.8	1.5	0.3	3.4	2.5	1.3	2.6	5.6	1.2	0.8	0.1	4.0
	Lombardy	1.2	0.3	1.0	3.6	6.2	2.3	8.3	2.5	50.0	0.9	0.5	4.2	2.5	1.0	2.8	5.1	1.4	1.3	0.1	4.6
	Marche	1.7	0.8	1.1	4.3	7.1	2.1	9.4	1.6	7.1	43.0	0.5	2.3	2.1	1.5	2.5	5.5	1.2	2.5	0.0	3.7
	Molise	1.4	1.0	1.6	11.0	4.3	0.9	16.7	1.0	8.1	0.8	32.9	1.5	3.8	1.9	2.9	4.0	1.7	1.3	0.1	3.0
	Piedmont	1.0	0.3	0.9	3.0	5.9	1.8	6.9	1.9	10.3	1.1	0.3	49.5	2.1	1.2	2.9	4.2	1.5	1.0	0.3	4.0
	Puglia	1.5	1.8	1.7	4.4	5.3	1.7	7.9	2.2	6.8	1.4	0.6	2.4	47.7	1.3	3.3	4.3	1.0	1.5	0.0	3.3
	Sardinia	1.3	0.4	1.2	3.1	5.5	1.6	8.9	2.4	6.9	1.1	0.5	2.9	2.1	49.4	3.3	5.0	0.8	0.7	0.0	3.0
	Sicily	1.1	0.3	2.9	4.6	4.5	1.4	7.3	1.2	6.0	1.1	0.4	2.3	2.3	1.2	54.4	4.5	0.5	1.2	0.0	2.7
	Tuscany	1.3	0.3	1.0	4.7	6.8	2.0	8.4	2.2	7.6	1.1	0.5	2.8	2.4	1.4	2.9	48.6	1.4	1.2	0.0	3.6
	Trentino Alto Adige	0.7	0.1	0.8	2.3	5.1	2.5	5.4	1.5	7.7	1.0	0.3	3.2	1.5	0.6	2.0	5.8	52.1	1.0	0.1	6.3
	Umbria	1.9	0.8	1.4	4.1	5.8	2.1	9.1	1.3	6.9	1.7	0.2	2.4	2.8	0.8	3.1	5.1	1.1	46.7	0.1	2.7
	Valle D'Aosta	0.0	0.6	0.0	2.8	8.9	1.1	6.1	3.4	15.1	1.7	0.0	22.3	2.2	0.0	1.1	3.9	1.1	2.2	24.0	3.4
	Veneto	1.4	0.2	0.6	2.8	7.4	3.6	7.2	1.6	9.8	1.2	0.3	3.3	1.8	1.0	2.3	4.1	2.3	0.9	0.1	48.3

The RBKF at field level

To conduct a stratification of the RBKF at field level, we use the subject categories (SCs) of the WoS classification schema. In particular, we assign each cited publication to the SC of the hosting source (journal, conference, book, etc.). A "full counting" approach is adopted, meaning that a publication published in a multi-category journal is fully assigned to each SC associated to the journal. The cited publications of our dataset are distributed over 246 SCs, in turn grouped in 13 scientific macro-areas.[vii]

As an example, Table 4 shows the value of the Tuscany RBKF in the SCs falling in the Biomedical research area. In this area, the "made in" Tuscany 2010-2012 publications generate altogether 17516 gains, 9421 of which extra-regional.

Vice versa, Tuscany earns 8492 gains from publications by the other Italian regions. The overall balance is therefore positive and equal to +929 units. By analysing data related to the single SCs, it can be noted that half of them show a positive balance, from a minimum of +33 in Medical laboratory technology to a max of +502 in Pharmacology & pharmacy; the remaining half show a nihil balance in Virology, and negative in six SCs (from -14 in Anatomy & morphology to -320 in Hematology).

Table 4: Tuscany RBKF in the WoS subject categories falling in Biomedical research

Subject category	Extra-regional gains generated (a)	Earned gains (b)	RBKF (a-b)
Allergy	108	123	-15
Anatomy & morphology	19	33	-14
Chemistry, medicinal	1391	977	+414
Hematology	483	803	-320
Immunology	1332	1079	+253
Infectious diseases	240	426	-186
Medical laboratory technology	153	120	+33
Medicine, research & experimental	498	650	-152
Oncology	1499	1465	+34
Pathology	484	254	+230
Pharmacology & pharmacy	1989	1487	+502
Radiology, nuclear medicine & medical imaging	760	576	+184
Toxicology	313	347	-34
Virology	152	152	0
Total	9421	8492	+929

When extending the analysis to all areas, SCs with a higher inclination to export (or import) of new knowledge may be identified. Table 5 reports the case of the first 10 SCs registering the lowest RBKF and the highest RBKF values for the Veneto region. SCs highly inclined to import, with a RBKF value ranging between -606 (Oncology) and -120 (Physics, multidisciplinary), are top of the list.

Actually, the prevailing presence of Bio-Med sciences SCs, with three SCs falling in Clinical Medicine and four in Biomedical Research, is quite evident.

Less evident is the disciplinary concentration in the lower section of the table, with the only thing deserving attention being the presence of two SCs falling in Biology at the very bottom.

Table 5: Subject categories with the highest and the lowest RBKF, for the Veneto region

Subject category	Area	Extra-regional gains generated (a)	Earned gains (b)	RBKF (a-b)
Oncology	Biomedical research	921	1527	-606
Pharmacology & pharmacy	Biomedical research	741	1056	-315
Gastroenterology & hepatology	Clinical medicine	796	1059	-263
Physics, particles & fields	Physics	751	980	-229
Cardiac & cardiovascular systems	Clinical medicine	739	964	-225
Genetics & heredity	Clinical medicine	266	470	-204
Radiology, nuclear medicine & medical imaging	Biomedical research	527	715	-188
Geochemistry & geophysics	Earth and space science	176	342	-166
Immunology	Biomedical research	861	1005	-144
Physics, multidisciplinary	Physics	191	311	-120
…	-	-	-	-
Electrochemistry	Chemistry	232	126	106
Energy & fuels	Physics	382	243	139
Rehabilitation	Clinical medicine	267	123	144
Agriculture, dairy & animal science	Biology	285	133	152
Hematology	Biomedical research	961	804	157
Surgery	Clinical medicine	962	721	241
Engineering, electrical & electronic	Engineering	794	510	284
Clinical neurology	Clinical medicine	1361	967	394
Biochemistry & molecular biology	Biology	1753	1230	523
Cell biology	Biology	1425	773	652

The analysis of knowledge flows may also be carried out on pairs of regions in order to identify the SCs with the highest surplus or deficit in the bilateral relations between the two regional research systems considered. Table 6, for example, reports the analysis on the flows between Latium and Piedmont for the SCs falling in Earth and space science. The reported value of balances shows surplus and deficit by SC, from the Latium perspective. Overall, Latium exports more than it imports from Piedmont (308 vs 306) mainly by virtue of the flows generated by publications in Environmental sciences, in Geochemistry & geophysics and in Meteorology & atmospheric sciences. In the opposite direction the Latium RBKF is negative in only four SCs, and mainly in Geosciences, multidisciplinary (-38) and in Geology (-27).

Table 7 shows the extension to all areas of the above mentioned analysis, with reference to the bi-directional flows between Lombardy and Emilia Romagna, in order to identify the SCs showing the greatest spread between knowledge inflows and outflows from one region to the other. Table 7 takes the Lombardy perspective in determining the RBKF deficit or surplus. It is the other way around from the Emilia Romagna perspective. The upper part of the table shows not so important differences between SCs with the greatest decifit for Lombardy, while there is an evident imbalance of flows in the lower part of the table, with a substantial surplus of flows from Lombardy to Emilia Romagna especially in Physics, particles & fields (432) and in Astronomy & astrophysics (296).

Table 6: The Latium - Piedmont RBKFs in the WoS subject categories of Earth and space science

Subject category	Gains from Latium to Piedmont (b)	Gains from Piedmont to Latium (a)	RBKF (a-b)
Geosciences, multidisciplinary	59	97	-38
Geology	7	34	-27
Paleontology	9	22	-13
Green & sustainable science & technology	2	10	-8
Geography, physical	18	19	-1
Oceanography	2	1	+1
Limnology	3	1	+2
Water resources	14	11	+3
Mineralogy	8	1	+7
Environmental studies	9	1	+8
Environmental sciences	83	63	+20
Geochemistry & geophysics	61	37	+24
Meteorology & atmospheric sciences	33	9	+24

Table 7: The Lombardy – Emilia Romagna RBKFs in the bottom and top 10 WoS subject categories per Lombardy RBKF deficit and surplus

Subject category	Area*	Gains from Lombardy to Emilia Rom.	Gains from Emilia Rom. to Lombardy	RBKF
Dentistry, oral surgery & medicine	7	62	146	-84
Microbiology	5	95	171	-76
Engineering, biomedical	9	116	187	-71
Obstetrics & gynecology	7	104	167	-63
Chemistry, multidisciplinary	3	195	256	-61
Orthopedics	7	77	134	-57
Reproductive biology	5	73	126	-53
Chemistry, medicinal	6	67	113	-46
Hematology	6	326	370	-44
Energy & fuels	2	35	72	-37
…	-	-	-	-
Environmental sciences	4	308	233	+75
Instruments & instrumentation	9	121	45	+76
Urology & nephrology	7	126	46	+80
Cell biology	5	278	187	+91
Neurosciences	7	556	463	+93
Clinical neurology	7	356	247	+109
Medicine, research & experimental	6	172	61	+111
Oncology	6	558	403	+155
Astronomy & astrophysics	2	573	277	+296
Physics, particles & fields	2	496	64	+432

** 1, Mathematics; 2, Physics; 3, Chemistry; 4, Earth and Space Sciences; 5, Biology; 6, Biomedical Research; 7, Clinical Medicine; 8, Psychology; 9, Engineering.*

Conclusions

This work has applied a new approach to measure knowledge spillovers, based on linkages between cited publications "made in" a given region and citing publications "made in" other regions of the same country. In doing so, we have been able to construct a regional balance of knowledge flows, at both overall and field level. Compared to previous literature, the recourse to publication citations rather than patent citations offer

more precise and robust results, as the order of magnitude of observations is much higher than that of patents.

While results cannot be generalized to other countries, some emerging evidences could be of general interest. There is a positive and strong correlation between the size of scientific output of a region and the value of its RBKF. Larger regions are better able to export new knowledge. At the same time we registered a positive correlation between the share of intraregional gains and the size of overall scientific output of a region. This can be due to the fact that in large regions it is likely to find large research laboratories/groups conducting research on topics of common interest.

The study can be easily replicated in other national contexts, and provide useful information to the policy maker: at aggregate level it allows to measure the share of intra- vs extra-regional knowledge spillovers generated by a region's research system, compared to other regions. At field level, the RBKF allows to pinpoint the subject fields with a higher propensity to export (or import) new knowledge to (from) other regions. Moreover, its application on pairs of regions allows to identify the field with the highest surplus or deficit in the bilateral relations between the two regional research systems considered. The possible longitudinal analysis of the RBKF could support the assessment of the effectiveness of research policies undertaken over time.

We warn the reader that all limitations and assumptions typical of bibliometric analyses apply.

A possible extension of the current work could measure the specialization indexes for outflows (export) and inflows (import) of knowledge by a given region. Such indexes measure respectively a region's capacity to "export" knowledge to other regions, or to "import" knowledge from other regions, as compared to the rest of the regions in the country, across all research fields. In simple terms, they measure the extent to which a region's knowledge flows differ from those of the rest of the regions or a comparison group of regions.

In addition to all possible methodological improvements, future research could investigate the geographic distance of knowledge spillovers and whether the geographic proximity effect tend to fade away over time.

References

Abramo, G., D'Angelo, C.A., Carloni, M. (2019). The balance of knowledge flows. *Journal of Informetrics,* 13(1), 1-9.

Abramo, G. (2018). Revisiting the scientometric conceptualization of impact and its measurement. *Journal of Informetrics,* 12(3), 590-597.

Abramo, G., D'Angelo, C.A. (2018). Who benefits from a country's scientific research? *Journal of Informetrics,* 12(1), 249-258.

Hassan, S.-U., & Haddawy, P. (2013). Measuring international knowledge flows and scholarly impact of scientific research. *Scientometrics*, 94(1), 163-179.

Jaffe A.B., Trajtenberg, M, &. Fogarty, S. (2000). Knowledge spillovers and patent citations: evidence from a survey of inventors. *American Economic Review*, 90(2), 215-218.

Jaffe A.B., Trajtenberg M., Henderson R. (1993). Geographic localization of knowledge spillovers as evidenced by patent citations. *The Quarterly Journal of Economics*, 108(3), 577-598.

Mehta, A., Rysman, M., & Simcoe T. (2010). Identifying the age profile of patent citations. *Social Science Research Network*, 25(7), 1179-1204.

Rabkin, Y. M., Eisemon, T. O., Lafitte-Houssat, J.-J., & McLean Rathgeber, E. (1979). Citation visibility of Africa's science. *Social Studies of Science*, 9(4), 499-506.

Romer, P. (1990). Endogenous technological change. *Journal of Political Economy*, XCVIII, S71-S102.
Romer, P. (1986). Increasing returns and long-run growth. *Journal of Political Economy*, XCIV, 1002-37.
Stegmann, J., & Grohmann, G. (2001). Citation rates, knowledge export and international visibility of dermatology journals listed and not listed in the Journal Citation Reports. *Scientometrics*, 50(3), 483-502.
Van Leeuwen, T. & Tijssen, R. (2000). Interdisciplinary dynamics of modern science: analysis of cross-disciplinary citation flows. *Research Evaluation*, 9(3), 183-187.

---

[i] We refer the reader to Abramo (2018) for a thorough discussion on the subject.
[ii] We include self-citing publications, because it may not matter whether the subsequent development that flows from a publication is performed by the same author(s), as long as it is performed in her or his region.
[iii] The spillovers we measure do not account for the sharing of knowledge among co-authors inherent in any research collaboration.
[iv] SCI-E: Science Citation Index Expanded; SSCI: Social Sciences Citation Index; A&HCI: Arts & Humanities Citation Index; CPCI-S: Conference Proceedings Citation Index- Science; CPCI-SSH: Conference Proceedings Citation Index- Social Science & Humanities; BKCI-S: Book Citation Index– Science; BKCI-SSH: Book Citation Index– Social Sciences & Humanities.
[v] Such publications are co-authored by an even number of Italian scholars, 50% of which affiliated to institutions located in the first region and 50% in the second.
[vi] The same as in column 8 of Table 1.
[vii] Mathematics; Physics; Chemistry; Earth and Space Sciences; Biology; Biomedical Research; Clinical Medicine; Psychology; Engineering; Economics; Law, political and social sciences; Art and Humanities; Multidisciplinary Sciences. The macro-areas and the assignment of SCs to them were at some point defined by the Institute of Scientific Information (ISI), although no longer showing in current Clarivate Analytics bibliographic products. There is no multi-assignment of SCs to macro-areas.

Can Altmetrics Supplement Citation Analysis for Funding Program Evaluation? Altmetric Analyses of National Cancer Institute (NCI) Extramural Divisions

Holly Wolcott[1], Duane Williams[1], Melissa Antman[2], James Corrigan[2], and Christine Burgess[1]

[1] *christine@uberresearch.com, holly@uberresearch.com, and duane@uberresearch.com*

625 Massachusetts Ave 2nd floor, Cambridge, MA 02139 (USA)

[2] *Melissa.Antman@nih.gov and Corrigan@mail.nih.gov*

Center for Research Strategy, National Cancer Institute, 31 Center Drive, MSC 2580 Building 31, Room 10A10 Bethesda, MD 20892-2580 (USA)

Abstract

This research-in-progress paper focuses on the use of altmetric indicators to gain insight into the influence of scientific publications linked to National Cancer Institute (NCI) funding. Within the National Institutes of Health (NIH), the NCI sponsors a diverse research portfolio ranging from basic research to advanced clinical applications and implementation science. This requires a diverse array of measures to accurately assess the impact of its programs and policies. To gain insight into the patterns of engagement with NCI research, we examined NCI's portfolio of extramural research, by NCI Division, and by research area. Patterns of altmetric indicators were observed to differ by research area. For example, the Division with the strongest public health focus (DCCPS) has the highest proportion of grants with Policy Document mentions. Additionally, the Division focused on the translation of promising research into clinical applications (DCTD) has a higher proportion of patent mentions. The online attention received by NCI Divisions was also compared to a reference set of biomedical publications to understand how differences in the subject area emphasis between NCI divisions influences altmetric counts.

Introduction

The United States National Cancer Institute (NCI) funds a diverse research portfolio. The comprehensive evaluation of such a diverse research portfolio needs to appropriately consider the impact of research across a broad range of stakeholders including scientists, clinical researchers, and the broader public. Citations are one measure of the extent to which research findings are visible to the research community, but citations only capture engagement of other scientists and often only in the same field. Alternative metrics, also known as "altmetrics", provide perspective on online research engagement of a broader set of communities including medical practitioners, patients, and the public (Loeb, 2014 & Kapp, 2015). Altmetrics provide insight into how the public engages with research on the web which allows funders to better understand how their research affects individuals outside of the traditional academy (Dinsmore, 2014 & Thelwall, 2016). Like citation-based metrics, altmetrics can vary based upon the research topic (Eysenback, 2011).

In this on-going research, we evaluate the altmetrics received by grants from four NCI extramural Divisions, listed below with their primary research emphasis:

1. Division of Cancer Biology (DCB): supports basic research in all areas of cancer biology.
2. Division of Cancer Control and Population Science (DCCPS): conducts and supports an integrated program of research to reduce risk, incidence, and death from cancer.
3. Division of Cancer Prevention (DCP): conducts and supports research to prevent and detect cancer.
4. Division of Cancer Treatment and Diagnosis (DCTD): supports the translation of promising research into clinical applications.

This analysis includes overall altmetrics, as well as four specific types of altmetrics (Twitter mentions, News Outlet mentions, Policy Document citations, and Patent citations). The four NCI Divisions examined here exhibit different altmetrics. In order to place the differences observed between NCI Divisions in context, we have also evaluated the altmetrics of NCI acknowledging publications in four Broad Research Areas (basic research, clinical medicine and science, health services research and public health) and compared this to a reference set of biomedical publications.

Methods

The publication and grant data utilized in this analysis were obtained from the Digital Science Dimensions platform. Altmetric data were obtained from Altmetric.com. A full list of the data sources tracked by Altmetric is available here: https://help.altmetric.com.

Grants included in this analysis had start dates between 2013 and 2017 and had NCI as the primary funder. In addition, only grants administered by one of the four extramural NCI divisions (DCP, DCCPS, DCP, DCTD) were included. Grants from the NCI Office of Director were excluded. The total number of grants included in this analysis as well as the total funding amount associated with these grants is provided in Table 1.

Table 1. Number of grants included in this analysis listed by NCI division, including total funding amount and the number of acknowledging publications.

NCI Division	*Number of Grants*	*Total Funding Amount (US$ Billions)*	*Total Publications*
DCTD	2,082	3.29	14,414
DCP	573	1.00	3,470
DCCPS	895	1.27	5,370
DCB	2,214	2.50	12,527

Two publication datasets were utilized in this study. The first dataset, "publications acknowledging NCI grants", includes publications between 2013 and 2018 that acknowledge one of the grants included in the grant dataset described above. The links between grants and publications originate from the Dimensions data. Dimensions links publications to grants using Funder provided data as well as analysis of the acknowledgement sections of full text publications for grant codes. The number of publications acknowledging NCI grants by NCI Division is provided in Table 1. A total of 28,527 publications are included in the NCI

acknowledging publications data set, and one publication may acknowledge funding from multiple NCI Divisions. A second reference publication dataset, "biomedical reference", includes all publications published in PubMed classified into a Broad Research Area (described below) between 2013 and 2018 where at least one author was located in the United States (n = 828,810). For both publication datasets, citations and Altmetric attention were obtained on 1/7/2019.

This analysis used the Broad Research Areas classifications scheme, which includes four categories: basic research, clinical medicine and science (clinical medicine), health services research (health sciences) and public health. Dimensions has implemented the Broad Research Areas developed by the Australian Bureau of Statistics (ABS) and published as part of the Australian and New Zealand Standard Research Classification (ANZSRC) 2008 edition, based on a machine learning approach. A training set has been used to build the matching model and to assign a code to all document types in Dimensions including grants and publications. The Broad Research Areas model classified 97% of grants and 100% of NCI acknowledging publications into Broad Research Areas.

Findings

The percentage of grants with publications, cited publications, and altmetrics was determined for the four NCI divisions (Table 2). DCTD had the highest percentage of grants with acknowledging publications (83.7%), while DCCPS had the lowest percentage of grants with acknowledging publications (69.6%). The amount of altmetrics varied between NCI Divisions, with DCB having the highest percentage of grants receiving online attention. As with citations, DCCPS had the lowest percentage of grants that have received online attention. Different distributions of funding mechanism and grant ages between the Divisions may be influencing the observed differences and will be investigated as part of on-going research.

Table 2. Percent of grants with publications, cited publications, and altmetrics for the four NCI Divisions.

NCI Division	*% Grants with Publications*	*% Grants with Cited Publications*	*% Grants with Altmetric Attention*
DCTD	83.7%	77.6%	76.2%
DCP	77.1%	71.2%	72.4%
DCCPS	69.6%	60.6%	62.9%
DCB	82.3%	77.2%	78.7%

Additional differences between NCI divisions were observed when four specific types of altmetrics (Twitter mentions, News Outlet mentions, Policy Document citations, and Patent citations) were examined in more detail (Table 3). For example, DCCPS grants received the most policy attention (6.9%) while also having the fewest number of grants that have received attention from Twitter, News Outlets, and Patents. DCTD grants have the highest patent activity, with 10.2% of grants acknowledged by one or more publications cited by a patent.

Table 3. Percent of grants with publications receiving Twitter attention, mentions from News Outlets, mentions in Policy Documents, and Patent citations.

NCI Division	*Twitter Accounts*	*News Outlets*	*Policy*	*Patents*
DCTD	75.5%	40.4%	3.7%	10.2%
DCP	71.6%	40.1%	5.8%	5.9%
DCCPS	62.0%	30.6%	6.9%	1.5%
DCB	78.3%	45.5%	1.4%	7.0%

In order to gain insight into the differences observed in the altmetrics for grants supported by each NCI Division, the subject area of the grant portfolio was determined. Table 4 shows the breakdown of each NCI Division's grants into Broad Research Areas. This classification scheme was selected because it differentiates between important aspects of biomedical research such as basic science and clinical medicine. DCB funded grants were primarily classified as Basic Science (90%), which is in alignment with the DCB's stated goal of facilitating basic research. Approximately half of DCTD and DCP grants are classified as clinical medicine, with a substantial part of their grant portfolio dedicated to basic science. The majority of DCCPS grants (62.1%) were classified as public health.

- **Table 4. Percent of Division Grants in the four Broad Research Areas**

NCI Division	Basic Science	Clinical Medicine	Health Services Research	Public Health	Unclassified
DCTD	47%	51.6%	0.4%	1.1%	0.0%
DCP	31%	52.7%	6.8%	9.4%	0.1%
DCCPS	8%	22.1%	7.2%	62.1%	0.1%
DCB	90%	9.6%	0.2%	0.5%	0.1%

A comparison between NCI acknowledging publications and a biomedical reference set was conducted to understand how differences in research emphasis between Divisions may impact the differences in altmetrics observed for the different Divisions. The biomedical reference set consisted of PubMed indexed publications with at least one US based author published between 2013 and 2018. PubMed indexed publications were selected as the reference set because these publications are representative of biomedical research. The comparison set was restricted to publications with at least one US based author because altmetrics are known to vary between countries (Alperin, 2013). Different levels of online attention across the full set of sources that Altmetric tracks were observed for the publications in the reference set (Table 5). For example, a higher percentage of public health articles received a Policy or News Outlet mention than publications in the other categories. A higher percentage of basic science publications received a patent citation when compared to the other categories

Table 5. Altmetrics for the biomedical reference publication set, by Broad Research Area.

Broad Research Area	N	% Cited	% Altmetric	% Twitter	% News	% Policy	% Patent
Basic Science	299,265	89%	71%	66%	12%	0.7%	1.67%
Clinical Medicine	385,248	81%	62%	58%	9%	1.8%	0.59%
Health Services	34,720	76%	69%	65%	9%	3.8%	0.03%
Public Health	119,158	81%	73%	68%	14%	5.0%	0.04%

An analysis of NCI acknowledging publications by Broad Research Area and the Division of the supporting grant is provided in Table 6. Note this analysis is at the publication level rather than the grant level. The percentage of publications funded by the NCI Divisions receiving online attention in any of the sources Altmetric tracks, as well as in Twitter mentions and News Outlet coverage specifically, equaled or exceeded the percentages observed for the biomedical research reference publication set. DCCPS exceeded the percentage of Policy Document mentions for the biomedical research reference publications set in basic science, clinical medicine, and health services.

Table 6. A comparison of NCI acknowledging publications by Broad Research Area and Division of acknowledged grant. Areas that met or exceeded the baseline calculated using the biomedical reference publications are highlighted in grey.

Broad Research Area	Division	Publication Count	% Cited	% Altmetric	% Twitter	% News	% Policy	% Patent
Basic Science	DCB	10597	85%	80%	77%	20%	0.2%	1.7%
	DCCPS	552	87%	86%	83%	28%	1.3%	2.0%
	DCP	1421	84%	73%	68%	14%	0.3%	1.3%
	DCTD	6475	83%	75%	72%	18%	0.2%	2.7%
	All	16661	85.6%	78.0%	74.6%	17.2%	0.3%	2.2%
Clinical Medicine	DCB	1518	85%	76%	73%	22%	0.5%	1.8%
	DCCPS	1180	82%	78%	75%	24%	1.1%	0.7%
	DCP	1669	85%	78%	75%	29%	2.3%	1.1%
	DCTD	7491	87%	76%	73%	25%	3.3%	1.8%
	All	8520	84.3%	73.3%	69.6%	18.4%	1.4%	1.3%
Health Service	DCB	0	--	--	--	--	--	--
	DCCPS	128	70%	70%	67%	13%	5.5%	0.0%
	DCP	67	67%	79%	79%	13%	4.5%	0.0%
	DCTD	16	50%	69%	69%	13%	0.0%	0.0%
	All	283	70.7%	77.4%	74.6%	15.5%	4.2%	0.0%
Public Health	DCB	53	77%	89%	87%	40%	1.9%	0.0%
	DCCPS	2320	81%	80%	75%	26%	3.2%	0.1%
	DCP	247	85%	83%	78%	30%	5.3%	0.4%
	DCTD	188	82%	83%	80%	29%	2.7%	0.0%
	All	3063	80.1%	77.8%	73.9%	23.5%	3.3%	0.1%

Discussion and Research in Progress

The preliminary analysis presented here suggests that altmetrics can potentially provide additional insight into the impact of funded work beyond that achieved through typical bibliometric analyses of the research literature based on publication and citation count data. The four NCI Divisions examined here exhibit different levels of attention across Twitter, News Outlets, Policy Documents, and Patents. These results suggest alignment of altmetric counts with the varied missions of the Divisions. For example, the Division with the strongest public health focus (DCCPS) has the highest proportion of grants with Policy mentions. Additionally, the Division focused on the translation of promising research into clinical applications (DCTD) has a higher proportion of grants with patent citations.

This analysis also highlights the importance of research's subject areas when interpreting altmetrics. Public Health research in the reference publication set has the highest level of

mentions in Policy Documents, which may play a role in the relatively high percentage of DCCPS grants with a Policy Document mention.

Further research will provide a deeper understanding of the Division-specific patterns of altmetrics and how to best use the observed baseline to understand the different altmetrics observed for the four Divisions. This includes conducting a more detailed analysis of how altmetrics vary with grant funding mechanism and funding level. In addition, we will consider if there are any factors at the Division level, such as a targeted social media presence, that influences the altmetrics.

Acknowledgments

The authors thank Stacy Konkiel (Digital Science) for her review of the abstract.

References

Alperin, J., (2015) Geographic variation in social media metrics: an analysis of Latin American journal articles. Aslib Journal of Information Management 67 289-304. doi: 10.1108/AJIM-12-2014-0176

Dinsmore, A., Allen, L., Dolby, K., (2014) Alternative Perspectives on Impact: The Potential of ALMs and Altmetrics to Inform Funders about Research Impact. PLOS Biology 12, e1002003. doi: 10.1371/journal.pbio.1002003

Eysenbach, G. (2011) Can Tweets Predict Citations? Metrics of Social Impact Based on Twitter and Correlation with Traditional Metrics of Scientific Impact. J Med Internet Res 13(4):e123. doi: 10.2196/jmir.2012

Kapp, J., Hensel, B., & Schnoring, K. (2015). Is Twitter a forum for disseminating research to health policy makers? Annals of Epidemiology 25: 883-887. doi: 10.1016/j.annepidem.2015.09.002

Loeb, S., Bayne, C., Frey, C., Davies, B., Averch, T., Woo, H., Stork, B., Cooperberg, M., & Eggener, S. (2013). Use of social media in urology: data from the American Urological Association (AUA). Surgical Education 113: 993-998. doi:10.1111/bju.12586

Thelwall, M., Kousha, K., Dinsmore, A., Dolby, K., (2016). Alternative metric indicators for funding scheme evaluations. Aslib Journal of Information Management 68, 2–18. doi: 10.1108/AJIM-09-2015-0146

Research of Competition Pattern and Technology Development Trend based on Patentometrics ——a Case Study of AI-field

Zhao Rongying[1] Li Xinlai[2] Li Danyang[3]

[1] *zhaorongying@126.com*
Research Center for China Science Evaluation, Wuhan University, Wuhan (China)
School of Information Management, Wuhan University, Wuhan (China)
Center for Studies of Information Resourses, Wuhan University, Wuhan (China)

[2] *lixinlai_whu@163.com*
Research Center for China Science Evaluation, Wuhan University, Wuhan (China)
School of Information Management, Wuhan University, Wuhan (China)
Center for Studies of Information Resourses, Wuhan University, Wuhan (China)

[3] *whusimldy@163.com*
Research Center for China Science Evaluation, Wuhan University, Wuhan (China)
School of Information Management, Wuhan University, Wuhan (China)
Center for Studies of Information Resourses, Wuhan University, Wuhan (China)

Abstract

In recent years, AI technology is drawing widespread attention around the world, so grasping the basic competition pattern accurately and identifying technology trends are of significance for governments, high-tech enterprises and even ordinary scholars. This paper studies global competition environment, technological competition pattern and technology trend by analysing the patent application, region distribution, quantity and quality of patentees, co-occurrence of IPC and MC. AI Technology is in the period of rapid growth, and China, Japan and the United States are the main countries applying for AI technology patents, with huge technical strength and market. The competitive enterprises including IBM, Panasonic, Microsoft, NEC Corporation, NTTPC, Canon, Toshiba, AT&T, Sony Corporation, and Mitsubishi. They have the same core technical categories on the date processing technology about speech or image recognition, but have different application due to the different AI development directions. The tendencies of AI technology includes mainly intelligent education, biological identification, intelligence server and intelligent terminal, big data analysis and information security. Some suggestions are put forward for the enterprises in the AI field at last.

Keywords

Patentometrics; Artificial intelligence (AI); Technological frontier; Competition pattern; Technology development trend; Co-occurrence analysis; Social network analysis

Introduction

In recent years, the artificial intelligence (AI) technology is drawing widespread attention around the world. Developed countries and regions have promoted AI developing as a national strategy, including the United States, the European Union, Japan and South Korea. Then improving the AI competitiveness continuously is the most important direction for the development of various countries. Therefore, grasping the basic competition pattern accurately and identifying technology trends of the AI are of significance for governments, high-tech enterprises and even ordinary scholars.

This paper studies the national competitive environment in the AI field by analysing the trend of patent application, distribution of regions and the highly cited patents. According to the quantity and quality of patentees and the IPC classification number, the main competitive enterprises in the AI-field and the core technologies of each enterprise are analysed. In addition, the research hotspots and technological frontier are identified through the co-occurrence analysis of Derwent manual code, so as to predict the future trend of technology development. The paper not only hopes to understand the development status and competition

pattern of international AI technology, but also hopes to provide references for the technical strategic layout of AI enterprises, especially in the future technology research and development direction.

Literature review

The technology competitiveness is the ability of enterprises gaining advantages in technological competition (Mayindi, 2010). In other word, it is the collection of many technical resources that are unique or impossibly imitated for other competitors (Wernerfelt, B, 1984). So the quantity, quality and efficiency of technological resources is the main factors for enterprises to obtain competitive advantage.

Patent literatures are the most effective carriers of technical information. It is even the only place where key technical information is disclosed for obtaining intellectual property somewhile (Fang et al., 2011). Besides, patent literatures, due to the advantages of easy access and uniform format, are analysed conveniently (Yoon B, 2008). Therefore, many scholars believe that patents are important information resources for analysing and evaluating the technological competitiveness of enterprises deeply. Also, they can be used to explore research hotspots and technology frontiers in a certain research field.

Kyebambe et al. (2017) propose a novel algorithm to automatically label data and then use the labeled data to train learners to forecast emerging technologies, which can retrieve as high as 70% of emerging technologies in a given year with high precision. Burhan et al. (2017) examined the patent filing behavior in PFROs by analyzing various motives that drive the patent filing of its researchers and Negative Binomial Regression Models were constructed to explain whether these patent filing motives impact patent portfolios in PFROs. Chen et al. (2012) constructed an evaluation index system of the enterprises' technological innovation ability by extracting the quantity index and quality index of patents from the patent literature. Then, based on the system, they evaluated the technological innovation ability of enterprises in the 3G technology field by using the close-value evaluation model. Zhou et al. (2014) studied the impact of patent citation on the technology competitiveness of RFID-related enterprises in the Chinese market, mainly from the direct citation and the total citation of patents. Li et al. (2014) obtained the three main components by factor analysis and principal component analysis at the company level, namely patent application, patent cooperation and patent technology. And they compared the technology competitiveness of companies in the OCT field using principal component synthesis model and three-dimensional patent portfolio analysis.

At present, the technology competitiveness of enterprise in the AI field is investigated seldom, and it is necessary increasingly. This paper analyses the patent application, the regional distribution and the IPC classification number in the AI field through methods related to patent map. We explore the global competitive patterns and core technologies of typical enterprises based on the results, as well as identify research hotspots and future development trends.

Data and Method

The data source of the paper is the Derwent Innovations Index (DII) published by Clarivate Analytics, part of Thomson Reuters. The database contains patent information from 47 patent offices around the world (covering more than 100 countries), with data tracing back to 1963 and updated every week which can guarantee the accuracy and authority of patent data therefore.

The paper adopted the retrieval strategies below: TI=("artificial intelligence*" OR " Deep learning *"OR "Natural language processing*" OR "Speech Recognition*" OR "Computer vision*" OR "Gesture control* "OR "smart robot*" OR "Video recognition*" OR "Voice

translation*" OR "Image Recognition*" OR "Machine learning*") and the total of 21,048 valid records were obtained finally. The retrieval strategies is determined by the trial-and-error based on the classification of the AI field of Venture Scanner (2017) (Analyst and technology powered research firm on emerging technology industry), as well as drawing lessons from some relevant literature (Zhang et al. 2018). The retrieval time is August 8, 2018 (database update on August 7, 2018).

Results and Discussion

Global Competition Environment of AI

This paper studies the national competitive environment in the AI field by analysing the trend of patent application, distribution of countries and the highly cited patents.

Overall Patent Application Trends

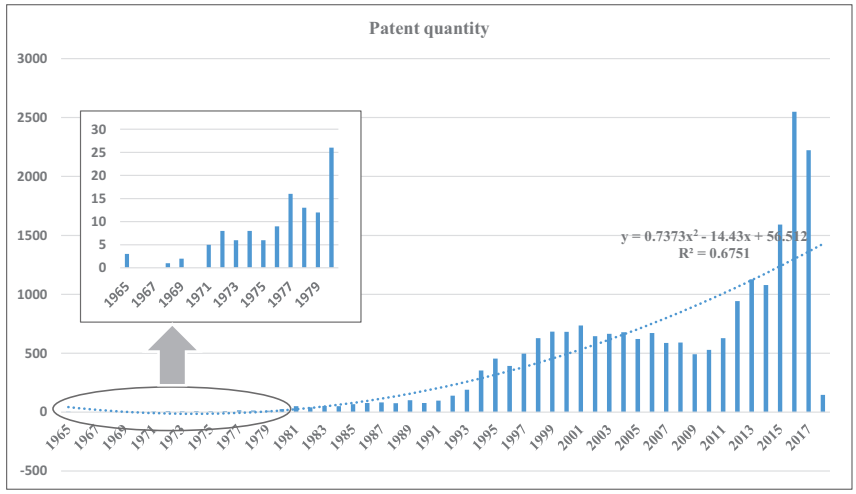


Fig. 1 Overall patent application trends

The development status of AI technology were investigated from the annual application of priority patents, as shown in Fig. 1 Overall patent application trends. Before 1989, the number of patent applications was no more than 100 per year, and the AI technology was in the embryonic stage. Then AI technology started to develop rapidly, and the first small peak emerged in 2001 with the annual application of 735. The development of AI technology experienced a slight stagnation in the following decade, however. Then after 2010, AI technology is booming and the number of patents increased rapidly, reaching 2,550 in 2016, which is dozens of times than the number before 1989. The result of curve regression analysis suggested the tendency of increasing continually in application number.

Country/Region Distribution of Patent Applications

The country/region analysis of patent applications reveals the global layout of the technology. Generally, only in the regions with intensive technology research or large market development potential, importance is attached to the patent application. The statistical results of main countries/regions on the AI applicants are shown in the Fig. 2. It is suggested that China, Japan and the United States are the main three countries for AI patents applying, accounting for more than 80%. Among them, China stands in the first place with the number

of 6,138 totally (28.67% of the total applications), including 383 patents from Taiwan and two patents from Hong Kong. Japan and the United States follows with 6,089 (28.44%) and 5,787 (27.03%), respectively. Therefore, China, Japan and the United States are the most active three in the field of AI technology innovation, with the huge AI market. In addition, there are more than 1000 applications in South Korea, and exceeded 200 in the Germany, the European Patent Office and the United Kingdom. Hence, AI technology is facing to a great development prospect, with enterprises actively distributing the global patents and entering the international market.

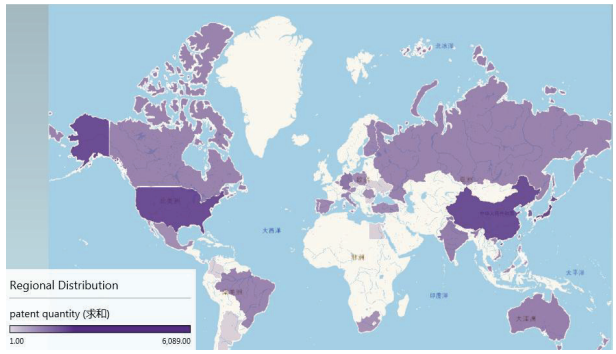


Fig. 2 region Distribution of patent applications

Analysis on Highly Cited Patents

Generally, the most highly cited patents, in a certain field or a certain period, occupy the key position which competitors cannot avoid in the industrial chain. The high citation frequency, therefore, reflects the fundamental and guiding role of the patent in its field. So the patent citation frequency is used to screen out the core and key patents sketchy.

The core patents in the AI field are explored by patent citation frequency, shown in Table 1. The citation frequencies of the patterns, at least 327, are far exceeding the average level in this field. It is believed that the top ten patents (patent numbers: US5963940-A, WO9621990-A2, US5799276-A, WO200065814-A1, EP376501-A, WO9118386-A, US6173066-B1, EP872827-A2, US6711293-B1, WO9919828-A1) are the core patents or contain the core technologies in the AI field. The continuous attention are paid to these patterns since they were applied. Also, the longest term of protection for invention patent is maintained. Base on the most highly cited patents, the technologies of natural language processing, speech recognition and computer vision are the fundamental and foremost part of the AI field.

Table 1 core patents in the field of AI (top 10)

Number	Title	Key Point	IPC	Time Of Application	Patentee (AE)	Cited	Time For Being Cited
1	Computerized information retrieval system operating method for natural language processing techniques	Natural language processing	US5963940-A	1995	UNIV SYRACUSE	677	1995-2015

2	Information system having speech interface and device for controlling access to broadcast data - has processor for executing speech recognition algorithm using received vocabulary data to recognise user speech instructions and control access to broadcast data	Speech recognition	WO9621990-A2; AU9647488-A; WO9621990-A3; US5774859-A	1995	SCIENTIFIC-ATLANTA INC	434	1995-2016
3	Knowledge based speech recognition system - selects word choice by comparing acoustic feature vector with acoustic event transcriptions of multiple words	Speech recognition	US5799276-A	1995	ACCENT INC	415	1995-2016
4	Dialog interaction defining method between speaker and speech recognition mechanism in interactive voice response system, by defining extensible class to instantiate class as object to control dialog interaction	Speech recognition	WO200065814-A1; AU200041854-A; US6314402-B1	1999	NUANCE COMMUNICATIONS	393	1999-2014
5	Speech recognition system - displays word menu to user for acceptance or prior editing before association with acoustic model	Speech recognition	EP376501-A; ES2018761-A; US5027406-A; EP376501-B1; DE68928097-E	1988	DRAGON SYST INC	393	1988-2016
6	Voice recognition and verification system over telephone network - stores speech recognition and verification data to identify digits of prompted caller password and to verify caller is authorised user	Speech recognition	WO9118386-A; AU9179958-A; US5125022-A; US5127043-A; EP528990-A1;	1990	VCS IND INC	376	1990-2016
7	Three=dimensional object recognizing method for computer vision and automated shape recognition involves determining whether portion of three=dimensional object is contained within zoned area of interest	Computer vision	US6173066-B1	1996	CYBERNET SYSTEMS CORP	356	1996-2016
8	Automated speech recognition service accessible by client over packet switching network operating - recognising received speech information by applying automatic speech recognition algorithm in accordance with grammar	Speech recognition	EP872827-A2; JP10333693-A; CA2228917-A; MX9802754-A1; US6078886-A;	1997	AT & T CORP	356	1992-2015
9	Scale invariant features identification method in computer vision system, involves dividing each pixel region, based on the pixel amplitude extrema and producing component subregion descriptors for each subregion	Computer vision	US6711293-B1	1999	UNIV BRITISH COLUMBIA	342	2000-2016
10	Background subtraction apparatus for a clean edged extraction of a foreground image e.g. for computer vision systems	Computer vision	WO9919828-A1; AU9910998-A; US6411744-B1; US2002172433-A1	1997	ELECTRIC PLANET INC	327	1997-2016

Nature language processing tops the list of the most highly cited patents. The research scope of natural language processing contains of all theories and methods that help effective communication between humans and computers using nature language, which is an important developing direction of computer science and artificial intelligence. The patent was successfully applied by Syracuse University in 1995. The novelty of this patent is embodied in the ability of accepting natural language queries and determining the query intent based on the utterance score. Besides, it can interact with the system and improve system interpretation by users.

Speech recognition is the most important approach to obtain information in the AI technology, accounting for 60% of the top ten highly cited patents. Voice interaction is the most mature and the fastest growing technology in the AI field currently. In particular, the development of

deep learning technology has raised the development of speech recognition, speech synthesis and natural language processing to a new level. Speech recognition technology of the highly cited patents mainly includes speech recognition for broadcast data access, knowledge-based speech recognition, interactive speech response, verification of telephone networks, and speech recognition services for switching network operations.

The other three patents are related to computer vision technology. Computer vision is a technology that uses cameras and computers to identify, track, and measure targets instead of human eyes. The three patents on computer vision technology are presented in the paper, namely three-dimensional object recognition technology, Scale invariant features identification technology, and Background subtraction apparatus for computer vision systems.

Technological Competition Pattern in the Field of Artificial Intelligence

According to the quantity and quality of patentees, the major competitors in the AI-field are be analysed and the institutions can be divided into four types. Then the core technical themes of representative enterprises are be excavated further by co-occurrence network of IPC.

Major Competitors in the Global Market

The analysis on the patentee reveals the patent owners and competitors, which predict the future competitive patterns. The top 10 companies with the most patents were studied, based on the data of 21,408 AI patents from the Derwent database, shown in Table 2. International business machines (IBM) ranks first with 765 patents, accounting for 3.57%. Then, Panasonic is the second, followed by Microsoft. In terms of patent proportion, monopoly enterprise are not yet in existence at present, but the possibility of their emergence cannot be ruled out with the continuous development of AI technology. Obviously, the top 10 enterprises are all from Japan or the United States, which indicates that the two countries have higher technological strength than the others. Now the advanced AI technology is also led by the two.

Table 2 major competitors in the global market

Patentee Code	Enterprise Name	Country	Patent Quantity	Proportion
IBMC-C	International Business Machines Corporation(IBM)	USA	765	3.57%
MATU-C	Panasonic	Japan	483	2.26%
MICT-C	Microsoft Corporation	USA	470	2.20%
NIDE-C	NEC Corporation	Japan	463	2.16%
NITE-C	Nippon Telegraph and Telephone Public Corporation(NTTPC)	Japan	426	1.99%
CANO-C	Canon	Japan	350	1.63%
TOKE-C	Toshiba	Japan	324	1.51%
AMTT-C	AT&T Inc.	USA	323	1.51%
SONY-C	Sony Corporation	Japan	304	1.42%
MITQ-C	Mitsubishi	Japan	295	1.38%

Patents are generally regarded as an important indicator of the technical strength of institutions, and the quantity and the quality are two factors to measure the competitive advantage of institutions. The number of patents held by the institution represents the technical output capacity. The combinatorial analysis on the patentee is demonstrated in the Fig. 1, which based on the idea of Boston matrix diagram of patent quality and patent quantity. The results reveal the major competitors in the AI field clearly. Generally, the technological

competition varies greatly in different institutions and can be divided into four types as shown in Fig. 3.

a) Preponderant Institutions. They are characterized by both high quantity and high quality of patent, such as IBMC-C and MICT-C.
b) Quantity-dominance Institutions. They are characterized by high quantity only. There are no appropriate examples in the AI field.
c) Quality-dominance Institutions. They usually master few but core patents and their technical strength could make great influence on the related fields, such as AMTT-C and MATU-C.
d) Developing Institutions. They own few patents and the patents are of low quality.

In the AI field, most institutions are classified as developing institutions and only few institutions have developed into the quality-dominance or preponderant ones. Therefore, the AI technology is still in the stage of rapid development and most of the institutions need to improve technical strength further.

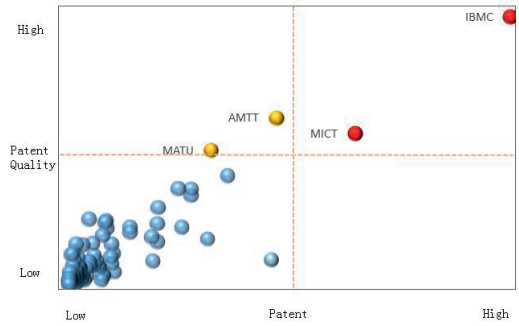


Fig. 3 institution matrix of patent quality and patent quantity

Core technical theme of Representative enterprises

Based on the classification of institutions in the AI-field, the paper uses the social network analysis method and clustering theory of Girven-Nerman algorithm to analyze the international patent classification number (IPC), so as to dig out the core technical theme of the Preponderant Institutions and Quality-dominance Institutions, namely IBMC-C, MICT-C, AMTT-C and MATU-C.

（1）IBMC-C (International Business Machines Corporation)

International Business Machines Corporation (IBM) has been a technology leader in the AI field since the 1950s. Recently, the development of AI is mainly implemented by IBM Watson, such as the AI cognitive services. According to the social network theory, the IPC co-occurrence shown in Fig. 4 suggests IBM's core technical theme. The first core theme is data processing for speech recognition and image recognition, including the IPC number G10L, G06F, G06K, G06T. The second core theme is computer systems based on specific computational models represented by IPC like G06N. The third core theme is telephonic communication represented by IPC like H04M.

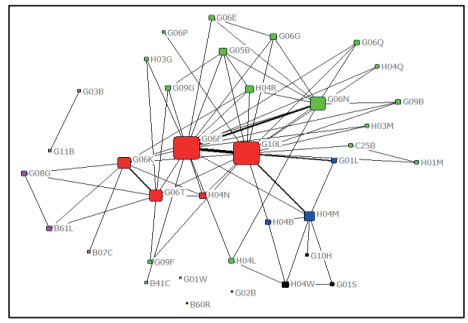


Fig. 4 IPC co-occurrence network of IBMC-C

（2）MICT-C (Microsoft)

Microsoft is also a pioneer in the AI field and has many well-known projects and products, such as Cortana, a built-in voice assistant in the Windows, and Zo, a chat robot. Meanwhile, Microsoft provides artificial intelligence services, such as robot services, machine learning and cognitive services. As shown in the Fig. 5, the core technology of Microsoft is focused on data processing for speech recognition and image recognition, including the IPC number G10L, G06F, G06K, G06T.

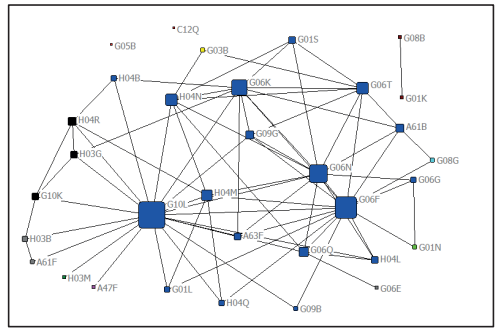


Fig. 5 IPC co-occurrence network of MICT-C

（3）AMTT-C（American Telephone & Telegraph Company）

American Telephone & Telegraph Company is the largest telecommunications company around the world. Therefore, different from the above two companies, AT&T's investigation on AI technology focuses on the communication mainly, such as 5G. The IPC co-occurrence shown in Fig. 6 suggests the AT&T's two core technical theme. The first core theme is electronic data processing for telephonic communication speech recognition, including the IPC number G10L, G06F, H04M. The second core theme is pictorial communication and gravitational measurements represented by IPC like H04N, G06K, G01V.

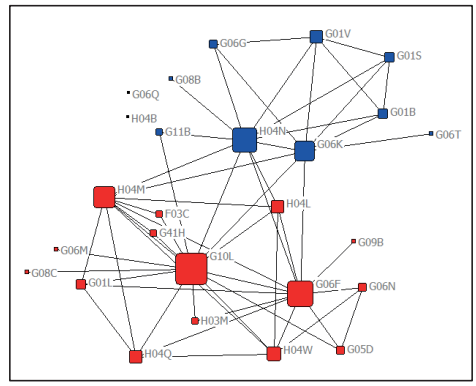


Fig. 6 IPC co-occurrence network of AMTT-C

（4）MATU-C (Panasonic Corporation)

Panasonic Corporation is a Japanese multinational electronics corporation. In recent years, Panasonic is also gradually combining AI technology with electrical products, such as intelligent air conditioners for air purifying automatically, and intelligent restaurants with robots serving. As shown in Fig. 7, Panasonic's core technology mainly focuses on the following aspects. The first core theme is electronic data processing for speech recognition telephonic communication, including the IPC number G10L, G06F, H04N, G06T. The second core theme is the technology of determining chemical or physical properties represented by IPC Numbers like G01N, G01B, H05K, etc. The third core theme is related to vehicles represented by IPC Numbers like B60K, B60R, E05B, E05F.

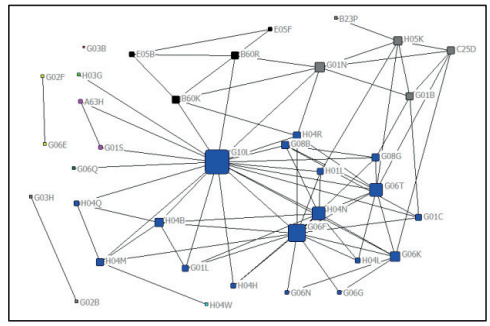


Fig. 7 IPC co-occurrence network of MATU-C

After the core technology analysis to the representative enterprises, it is suggested that the common ground in the AI field is the date processing technology about speech or image recognition with the IPC number like G10L, G06F, G06K, G06T, H04N, H04M. However, the core technology themes various in different enterprises due to the difference on the AI developing direction. The direction of development depends on the experience and history of the enterprise. IBM grew up with computer technology and Microsoft started by developing the operating system for personal computer, which lead the two enterprises to focus on the software and computer systems in the AI development. AT&T is telecommunications service provider, then it devotes more attention to the application of AI technology in mobile

communication. Panasonic, as an international electronic technology enterprise, invests more in the combining between AI technology and electronic products, aiming to create a smart life.

Development Trend of Patent Technology in AI Field

Through the Co-occurrence analysis of Derwent manual code, the research hotspots and technological frontier, so as to predict the future trend of technology development.

Research Hotspots

Co-occurrence analysis is an important research method in the scientometrics, and its application in patent analysis can play an important role in identifying research hotspots, revealing technological development trend, and describing industrial technology network structure. Knowledge map of patent classification codes is an excellent presentation of hot category and development trend in a period of time. The paper analyses the Derwent Manual Code (MC) of the past five years, and the category codes with high occurrence frequency indicates research hotspots. Fig. 8 is drawn by CiteSpaceV based on patents information during 2013 to 2018, where the size of the circle represents the frequency. Form the results, the co-occurrence frequency and centrality of MC can be understood clearly.

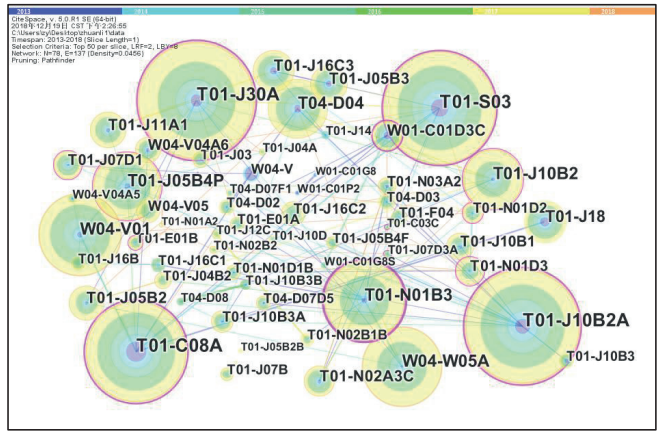


Fig. 8 co-occurrence network of Derwent manual code

Table 3 the top 10 technology category

Frequency	centrality	year	manual code(MC)	technology category
1521	0.21	2013	T01-J30A	Multimedia computer systems
1504	0.38	2013	T01-J10B2A	Image analysis
1467	0.33	2013	T01-S03	Claimed software products
1316	0.3	2013	T01-C08A	Speech recognition/synthesis input/output
1092	0.09	2013	W04-V01	Novel aspects of analysis or recognition
1054	0.03	2013	W04-W05A	Educational equipment in general
1047	0.43	2013	T01-N01B3	On-line education
903	0.17	2013	T01-J05B4P	Database applications
802	0.12	2013	T01-J10B2	Image analysis
789	0.09	2013	T04-D04	Recognition
569	0	2013	T01-J18	Computer processing for speech/audio

The top 10 MC with high occurrence frequency are shown in Table 3. Combined with Fig. 8 and Table 3, it can be seen that research hotspots are distributed in speech recognition, image recognition, application programs, application software, and intelligent education, which are consistent with the core technologies of the representative enterprises mentioned above. In addition, recent technological hotspots tend to focus on the research of biological identification, such as fingerprint identification, acoustic analysis, which is conducive to promoting the convenience and security of authorization and adapted to the current Internet development. In the aspect of speech recognition, research hotspots tend to semantic recognition in order to break through the "bottleneck".

Technological Frontier

Word frequency detection technology provided by CiteSpace V, by analyzing the time distribution of word frequency, can detect burst term from the large number of words. The variation of the frequency can be used to determine the frontier and developing trend. Therefore, it is believed that patent categories with large burst value represents the technology frontier.

Table 4 Subject Categories with the Strongest Citation Bursts (Top 10)

Subject Categories	technology category	Year	Strength	Begin	End	2013 - 2018
W04-V	Analysis, synthesis and processing of sound waves	2013	93.3038	2013	2014	
T01-J14	Language translation	2013	33.211	2013	2015	
W01-C01P2	Personal digital assistant	2013	34.4335	2013	2014	
T01-J10B3	Object processing	2013	13.6389	2013	2014	
T01-J10B3B	Object colour processing and colour system conversion	2013	32.9056	2014	2015	
T04-D08	Colour systems	2013	30.7612	2014	2015	
T04-D07D5	Detecting position or orientation	2013	32.6965	2015	2016	
T01-E01B	Sorting, selecting, merging or comparing data	2013	4.8321	2016	2018	
T01-J04B2	Correlation function	2013	27.2128	2016	2018	
T01-J12C	Security	2013	19.0554	2016	2018	

Table 4 shows that AI of biological identification (W04-V, T01-J10B3) is the key technologies, which have huge market prospects. As released by “MIT Technology Review”, “Paying with Your Face” is one of the top ten breakthrough technologies in the world in 2017. With the development of the Internet and the popularity of smart phones, application of AI server and intelligent terminal like language translation (T01-J14) and detecting position or orientation services (T04-D07D5) are also become the technological frontier at present. AI technology also shows great advantage in the big data management. The deep learning and machine learning are important increasingly for data management (T01-E01B), data analysis (W01-C01P2), information sharing (T01-J04B2), information security (T01-J12C) and so on.

Conclusion and suggestions

This paper studies the national competitive environment in the AI field by analysing the trend of patent application, distribution of countries and the highly cited patents. AI Technology is in the period of rapid growth, and it has huge development space and potential. China, Japan and the United States are the main countries applying for artificial intelligence technology

patents, with huge technical strength and market. Natural language processing, speech recognition and computer vision can be found as part of the basic and core technologies in the AI field.

According to the quantity and quality of patentees and the IPC classification number, the competitive enterprises are IBM, Panasonic, Microsoft, NEC Corporation, NTTPC, Canon, Toshiba, AT&T, Sony Corporation, Mitsubishi mainly in the AI field. The four representative enterprises have the same core technical categories on the date processing technology about speech or image recognition, but they have different application due to the different AI development directions.

In addition, the research hotspots and technological frontier are discovered to predict the AI development trend through the Co-occurrence analysis of DC. The tendencies includes mainly intelligent education, biological identification, AI server and intelligent terminal, big data analysis and information security.

Based on the careful AI industry analysis, the following suggestions are put forward for the enterprises in the AI field: (a) improve the quality of patents actively and increase the quantity of patents, aiming to become the Preponderant Institutions; (b) formulate the development strategy based on the enterprise's own resources and ability advantages, keeping up with the development trend of industry commercialization and synergism; (c) focus on introducing innovative talents and cultivate the core competence based on talent advantage, paying more attention to talents in the fields of biological identification, intelligent education, intelligent terminal and server software application.

Acknowledgments

This paper is supported by Major Project of National Social Science Foundation (Grant No. 18ZDA325) and National Social Science Foundation in China (Grant No.16BTQ055).

References

Burhan, M. , Singh, A. K. , & Jain, S. K. . (2017). Patents as proxy for measuring innovations: a case of changing patent filing behavior in indian public funded research organizations. Technological Forecasting & Social Change, 123.

Chen L. X., Zhao X.Q., &Zheng L. Y. (2012). Evaluating Firms' Technological Innovation Capability by Patent Information Analysis: the Case of Firms in the Communication Industry. Science & Technology and Economy. 25(1):37-41.

Fang,S., Zheng Y., Pang H. S. &Zhang X. (2011) Study on the Method of Analyzing Technology Evolution Based on Patent Documents. Library and Information Service. 55(22):42-46.

Kyebambe, M. N., Cheng, G. , Huang, Y. , He, C. , & Zhang, Z. . (2017). Forecasting emerging technologies: a supervised learning approach through patent analysis. Technological Forecasting and Social Change, S0040162516307065.

Li S. Y., Fang S. (2014). An Empirical Research of Patent Portfolio Analysis Method on the Company Level [J]. Journal of Intelligence. 33(3):39-43.

Mayindi, Kachienga. (2010). Technological competence and capacity building within firms: South Africa's civil aircraft industry// Technology Management for Global Economic Growth. IEEE.

MIT Technology Review. (2017). top ten breakthrough technologies in the world. Retrieved October 18, 2018 from http://tech.163.com/17/0221/15/CDQETUQT00098J1S.html

Shen J., Gao J. P., Teng L. (2012). Derwent Manual Code Co-Occurrence: A Practical Method in Patent Map. SCIENCE OF SCIENCE AND MANAGEMENT OF S. & T. 33(1):12-16.

Venture Scanner.(2017). 13 segments of Artificial intelligence. Retrieved October 15, 2018 from http://www.solves.com.cn/doc-view-4876.html

Wernerfelt, B. (1984) A Resource-Based View of the Firm. Strategic Management Journal, 5, 171-180.

Yoon B , Lee S . (2008) Patent analysis for technology forecasting: Sector-specific applications[C]// IEEE International Engineering Management Conference. IEEE.

Can the Emergence of New Developments in the Techno-Sciences be Indicated as "Hot Spots" in Journal Maps?

Xiaozan LYU,*[1] Ping ZHOU,[2] and Loet Leydesdorff[3]

[1] lvxz1991@zju.edu.cn ; [2] pingzhou@zju.edu.cn
Zhejiang University, Department of Information Resources Management, School of Public Affairs, Hangzhou, (China);
[3] loet@leydesdorff.net
Amsterdam School of Communication Research, University of Amsterdam, PO Box 15793, 1001 NG Amsterdam, (The Netherlands)

Abstract
Most documents in the Science Citation Index (Web of Science) contains a field with cited references (CR). One can measure the patent-intensity [N(PatRef)] of a paper or journal in terms of the absolute number of references to patents among these CRs, and the patent-density [%(PatRef)] as a percentage over total number of references [N(Ref)]: %(PatRef)= 100 * N(PatRef)/N(Ref). New technology-oriented developments can be expected to lead to increases in patent-intensity in the fields affected. Both the patent-intensity [N(PatRef)] and patent-density [%(PatRef)] can be used to color or shade a journal map and/or a map based on Web-of-Science Subject Categories (WCs) for a series of years. The resulting maps can be used as base-maps for overlaying portfolios. The envisaged software will enable users to generate the overlays for any set of documents from WoS.

Introduction

The study of the relations between scholarly publications and patent documents as output of R&D has focused one-sidedly on "non-patent literature references" (NPLR) in patents. In a number of publications, Narin and his colleagues studied the linkages between science and technology in terms of NPLR (e.g., Narin & Olivastro, 1988 and 1992). These authors noted an increasing linkage between science and technology in terms of mutual citation relations (Grupp & Schmoch, 1992). Glänzel & Meyer (2003) added that one can also study reversely the referencing of patents in scholarly publications.

The overlay techniques developed during the latest ten years (Leydesdorff & Rafols, 2009) enable us to visualize patent density among journal citations or *vice versa* NPLR in patent portfolios overlaid on base maps. Can this global and statistical approach inform strategic decision-making on emerging technologies (Porter, Roessner, Jin, & Newman, 2002; Rotolo, Rafols, Hopkins, & Leydesdorff, 2017)? In this study, we do not focus on the identification of the micro-emergence of topics in science-based technologies (Small, Boyack, & Klavans, 2014), but in the strategic overview based on changes in the information streams and accordingly portfolios (Kogler, Heimeriks, & Leydesdorff, 2018).

The study was initiated by a contest for indicators of emergence organized by the team at Georgia Tech under the aegis of the US-NSF (Carley, Newman, Porter, & Garner, 2018). We generate a base map at the journal level (Leydesdorff & Rafols, 2012), using a recursive algorithm for the field delineations among journals (Leydesdorff, Bornmann, & Wagner, 2017). The overlay option of VOSviewer is used for the visualization (Van Eck & Waltman, 2010). In a follow-up study we envisage to use a similar technique for generating patent portfolio maps of NPLR in USPTO data for the same period and then study the mutual interactions (Leydesdorff, Alkemade, Heimeriks, & Hoekstra, 2015; Leydesdorff, Kogler, &

Yan, 2017; Yan & Luo, 2017). Such a two-side comparison may help us answer the following questions: Are there any connections between the "hot spots" on the two maps? If any, to what extent do the "hot spots" on both maps correspond?

Methodology

We use the combined set of 11,679 journals listed in the Journal Citation Reports (JCR) 2017 of the Science Citation Index and the Social Sciences Citation Index. Using dedicated software, the data is first organized into a journal-journal citation matrix using the .net format of Pajek. VOSviewer can read this file; it symmetrizes the asymmetrical matrix internally by summing the cells (i, j) and (j, i). Two journals were not connected into the largest component (*Dutch Crossing* and *Z Bibl Bibl*), so that we work with (11,679–2 =) 11,677 journals.

Figure 1 provides the base map; Table 1 shows the statistics of the clustering at the first level of the hierarchy. Eleven broad fields are distinguished in total. The clustering at the specialty level of the map can be available at http://www.leydesdorff.net/jcr17/page1.htm.

Table 1: Clusters and numbers of journals

Cluster_id	Clusters	N of Journals	Colour in map
1	Social Sciences	3,385	**Red**
2	Clinical Medicine	2,240	**dark green**
3	Computer Science	1,814	**dark blue**
4	Environmental Sciences	1,704	**Yellow**
5	Nano Science & Technology	835	**Purple**
6	Biology	588	**light blue**
7	Chemistry	493	**Orange**
8	Neuro Sciences	301	**Brown**
9	Molecular	199	**light pink**
10	Ophthalmology	60	**Pink**
11	Astronomy	58	**light green**
	Grand Total	**11,677**	

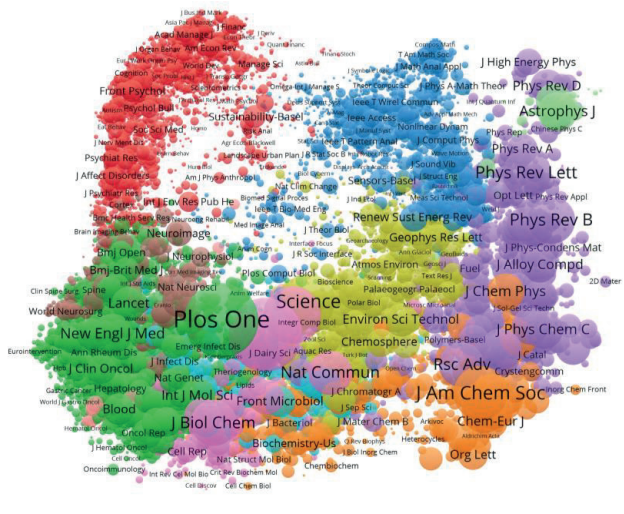


Figure 1: Eleven clusters of 11,677 journals belonging to the largest component of the JCR 2017 journal citation matrix; VOSviewer is used for classification and visualization. This map can be web-started at https://www.leydesdorff.net/jcr17/

A comprehensive list of identification numbers of scholarly documents citing patents during 2017 was obtained from the in-house database of the Center for Science and Technology Studies (CWTS) in Leiden. On this basis, we harvested 23,277 documents ("Articles", "Reviews" and "Letters") published in 3,207 (that is, 27.5% of the) journals. These papers contain 1,230,690 references of which 50,981 to patents. We distinguish USPTO patents as a subclass of patents; these are referenced 21,400 times by 2,473 documents.

Table 2 shows that papers with references to patents contain on average significantly more citations than average papers (52.62 and 33.21, respectively). Within the set of patent-referencing documents, only 4.41% of the references are to patents. For US patents, various parameters are smaller by an order of magnitude.

Table 2: Descriptive statistics of the numbers of journals, documents, and references

	JCRs 2017	Sample containing patent refs.	Refs. to Patents	Refs. to US patents
N of journals	11,679	3,207		2,473
N of documents	1,597,777	23,387		21,203
N references	(53,062,285)[a]	1,230,690	50,981	21,400
	64,296,601[b]	(2.32%)	(4.41%)	(1.74%)
References/Document	33.21	52.62	2.18	1.01

Notes: [a] This number is based on the citation counts in the citation matrix, while [b] indicates the number provided for the database by ISI/Clarivate.

Results

Figure 2 display the overlaid map based on number of cited patents in journals. From the top-left (Social Sciences) to the lower right side (Chemistry; Nano Science & Technology) the patent intensity is increasing. Are the sciences perhaps organized along a technological gradient? (cf. Narin, 2013; Price, 1984).

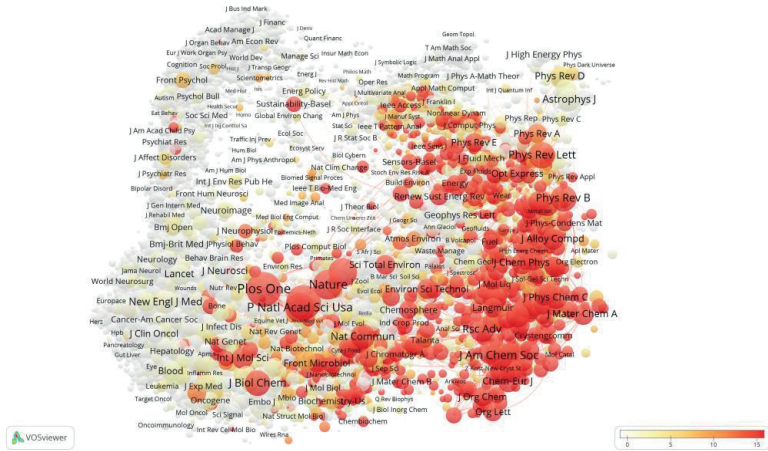
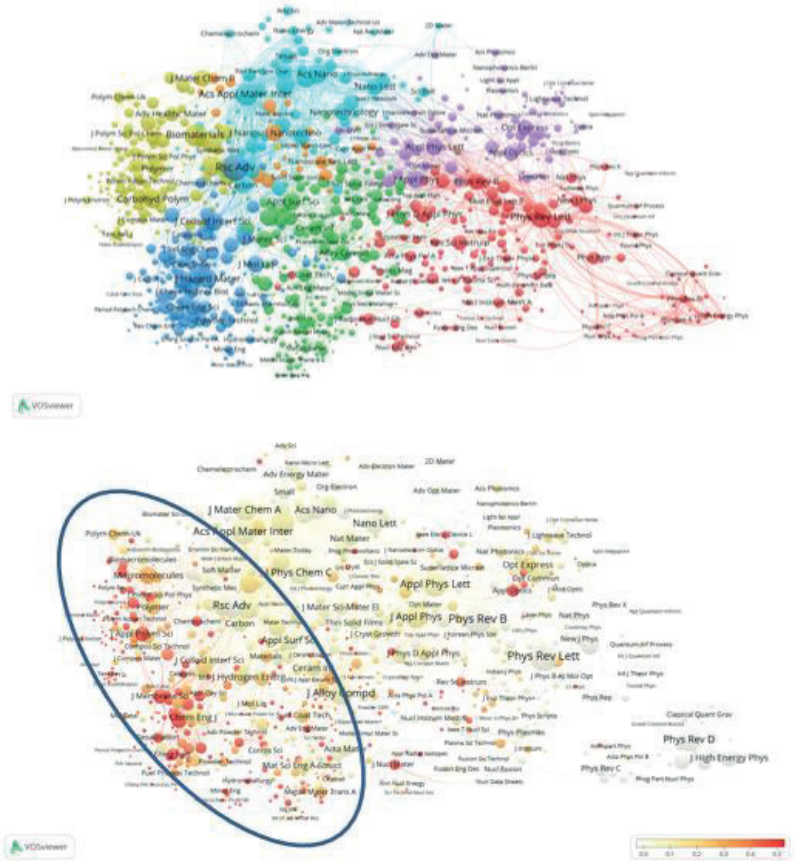


Figure 2: Numbers of patents cited in journals. This map can be webstarted at https://leydesdorff.net/jcr17/index1.htm .

In Figures 3A and B, we compare the base map and the patent-overlaid map for the 693 nano-journals organized into this fifth cluster. These maps also suggest that patent-referencing is concentrated in specific parts of the map. The maps based on percentages contain more shades (representing fine structure) than the ones based on patent numbers (Fig. 2).

References to patents in this sub-domain of Nano Science & Technology seem to be concentrated in a belt from chemistry and materials-related journals at the left with a gradient to physics journals at the right side. This belt represents the interface between chemistry and more fundamental nano-oriented sciences. The common denominator is probably instrumentation.



Figures 3A and B: Base Map and Overlay Map of 17,955 patents (expressed as percentages) in 693 nano-relevant journals organized into eight disciplinary and specialist clusters.

Discussion

Both the *patent-intensity [N(PatRef)]* and *patent-density [%(PatRef)]* can be used to measure the patent-citations at the aggregated cluster level. Technically, for each cluster, we calculate its patent-intensity and density in two ways: the arithmetic means of *N(PatRef)s* and *%(PatRef)s* of journals based on total number in the respective sample (i.e., journals citing patents). Table 3 displays the statistical description of clusters. As is shown,

more than 80% of the journals in Nano Science & Technology and in chemistry contain papers citing patents. Patent citations are numerically most frequent in chemistry journals, but the *patent-density [%(PatRef)]* in Nano Science & Technology is much higher than in Chemistry. This result points to the possible relevance of the density indicator for indicating "hot topics". In the older field of Chemistry citation of patents is common, but as a relatively new field, nano-related papers have a significantly higher percentage of patents referenced than those in Chemistry. Furthermore , the introduction and application of technologies is not always the reason for citing patents. The high preference for patent information of some Social Sciences journals proves that patenting is also an important objective in researches relevant to scientific evaluation or intellectual property.

Table 3: Statistical description of clusters

Clusters	N of journals In JCR	N of journals citing patents	(%)	%(PatRef)	N(PatRef)/ N(doc)	%(PatRef)	N(PatRef)/ N(doc)
	(a)	(b)	(c)	JCR domain (column a)		Respective samples (b)	
1. Social Sciences	3,385	141	4.17	0.01	0.16	**0.30**	**3.77**
2. Clinical Medicine	2,240	402	17.95	0.02	0.79	0.09	4.41
3. Computer Science	1,814	740	40.79	0.18	4.22	0.45	10.35
4. Environmental Science	1,704	404	23.71	0.06	1.61	0.26	6.79
5. Nano Science & Technology	835	693	**82.99**	**0.48**	**21.50**	**0.57**	**25.91**
6. Biology	588	262	44.56	0.09	4.90	0.19	10.99
7. Chemistry	493	401	**81.34**	**0.40**	**33.18**	**0.49**	**40.79**
8. Neuro Sciences	301	60	19.93	0.01	0.64	0.07	3.22
9. Molecular	199	73	36.68	0.40	3.75	1.10	10.22
10. Ophthalmology	60	19	31.67	0.05	1.98	0.15	6.26
11. Astronomy	58	12	20.69	0.02	0.41	0.10	2.00
Grant Total	**11,677**	**3,207**	**27.46**	**0.11**	**4.37**	**0.39**	**15.90**

We intend (*i*) to repeat the analysis for the period from 2013 to 2017, and then to animate the "hot spots;" and (*ii*) to complement the analysis with mapping NPLRs overlaid on patent maps in terms of the Collective Patent Classification (CPC) at the four-digit level. We envisage to analyze the "hot spots" as emerging developments using the tests for discontinuities developed by Leydesdorff, Wagner, & Bornmann (2018).

Acknowledgments

XL and PZ acknowledge support by the National Science Foundation of China (NSFC), Grant Numbers 71473219, 71843012. We are grateful to Clarivate Analytics for providing the JCR data and CWTS for the UT numbers of articles with paten citations. We thank Bowen Yan for delivering the patent data, which will be used in the next stage of the project.

References

Carley, S. F., Newman, N. C., Porter, A. L., & Garner, J. G. (2018). An indicator of technical emergence. *Scientometrics, 115*(1), 35-49.

Glänzel, W., & Meyer, M. (2003). Patents cited in the scientific literature: An exploratory study of "reverse" citation relations. *Scientometrics, 58*(2), 415-428.

Grupp, H., & Schmoch, U. (1992). Perceptions of scientification of innovation as measured by referencing between patents and papers: dynamics in science-based fields of technology. In H. Grupp (Ed.), *Dynamics of science-based innovation* (pp. 73-128): Springer.

Kogler, D. F., Heimeriks, G., & Leydesdorff, L. (2018). Patent Portfolio Analysis of Cities: Statistics and Maps of Technological Inventiveness. *European Planning Studies, 26*(11), 2256-2278. doi: 10.1007/s11192-018-2682-1

Leydesdorff, L., & Rafols, I. (2009). A Global Map of Science Based on the ISI Subject Categories. *Journal of the American Society for Information Science and Technology, 60*(2), 348-362. doi: 10.1002/asi.20967

Leydesdorff, L., & Rafols, I. (2012). Interactive Overlays: A New Method for Generating Global Journal Maps from Web-of-Science Data. *Journal of Informetrics, 6*(3), 318-332.

Leydesdorff, L., Alkemade, F., Heimeriks, G., & Hoekstra, R. (2015). Patents as instruments for exploring innovation dynamics: geographic and technological perspectives on "photovoltaic cells". *Scientometrics, 102*(1), 629-651. doi: 10.1007/s11192-014-1447-8

Leydesdorff, L., Bornmann, L., & Wagner, C. S. (2017). Generating Clustered Journal Maps: An Automated System for Hierarchical Classification. *Scientometrics, 110*(3), 1601-1614. doi: 10.1007/s11192-016-2226-5

Leydesdorff, L., Kogler, D. F., & Yan, B. (2017). Mapping Patent Classifications: Portfolio and Statistical Analysis, and the Comparison of Strengths and Weaknesses. *Scientometrics, 112*(3), 1573-1591.

Leydesdorff, L., Wagner, C., & Bornmann, L. (2018). Discontinuities in Citation Relations among Journals: Self-organized Criticality as a Model of Scientific Revolutions and Change. *Scientometrics, 116*(1), 623-644. doi: 10.1007/s11192-018-2734-6

Narin, F. (2013). Tracing the Paths from Basic Research to Economic Impact. *F&M Scientist 1*(1), 67-88.

Narin, F., & Olivastro, D. (1988). Technology Indicators Based on Patents and Patent Citations. In A. F. J. v. Raan (Ed.), *Handbook of Quantitative Studies of Science and Technology* (pp. 465-507). Amsterdam: Elsevier.

Narin, F., & Olivastro, D. (1992). Status Report: Linkage beteen technology and science. *Research Policy, 21*, 237-249.

Porter, A. L., Roessner, J. D., Jin, X.-Y., & Newman, N. C. (2002). Measuring national "emerging technology"capabilities. *Science and Public Policy, 29*(3), 189-200.

Price, D. J. de Solla (1984). The science/technology relationship, the craft of experimental science, and policy for the improvement of high technology innovation. *Research Policy, 13*(1), 3-20.

Rotolo, D., Rafols, I., Hopkins, M. M., & Leydesdorff, L. (2017). Strategic Intelligence on Emerging Technologies: Scientometric Overlay Mapping. [Article]. *Journal of the Association for Information Science and Technology, 68*(1), 214-233. doi: 10.1002/asi.23631

Small, H., Boyack, K. W., & Klavans, R. (2014). Identifying emerging topics in science and technology. *Research Policy, 43*(8), 1450-1467.

Van Eck, N. J., & Waltman, L. (2010). Software survey: VOSviewer, a computer program for bibliometric mapping. *Scientometrics, 84*(2), 523-538.

Yan, B., & Luo, J. (2017). Measuring technological distance for patent mapping. *Journal of the Association for Information Science and Technology, 68*(2), 423-437.

Allocation of non-competitive research funding to single researchers: preliminary analysis of the short-term effects

Domenico Maisano[1], Luca Mastrogiacomo[2] and Fiorenzo Franceschini[3]

[1] *domenico.maisano@polito.it* [2] *luca.mastrogiacomo@polito.it* [3] *fiorenzo.franceschini@polito.it*
Politecnico di Torino, DIGEP (Department of Management and Production Engineering),
Corso Duca degli Abruzzi 24, 10129, Torino (Italy)

Abstract

Research funding is essential to promote the scientific activity of researchers. Simplifying, funding schemes can be classified in two categories – *competitive* and *non-competitive* – with several corresponding advantages and shortcomings, which are widely discussed in the scientific literature.

The researchers of Politecnico di Torino (i.e., one of the major Italian technical universities) have recently been funded through a non-competitive research funding, consisting of 14k€ for every single researcher in each of the last three years (i.e., 2017, 2018 and 2019), for a total of 42k€.

This initiative – also called "diffused funding" (DF) – represents an important opportunity to investigate the effects of the relatively large allocation of non-competitive funding to single researchers. In this regard, this paper investigates the effects of the DF on the researchers' scientific output, according to four dimensions of analysis: *publishing productivity*, *publishing diffusion/impact*, *journal reputation*, and *international research relations*. Preliminary results do not reveal any improvement in the publication output, at least in the short term.

Introduction and literature review

There is no doubt that (public and private) funding is essential to encourage the constant and effective evolution of the scientific research and the dissemination of its results (Jacob and Lefgren, 2011). Different funding schemes can result from the combination of (at least) four factors:

1. *Funding body.* We distinguish between *non-commercial* – e.g., public government bodies, research councils, European commissions, etc. – and *commercial* – e.g., private companies or industrial consortia/districts, which usually propose research topics.
2. *Funding recipients.* We distinguish between individual researchers and research institutions – such as entire universities, departments, groups of researchers, etc. – who generally have some autonomy in allocating funds internally.
3. *Allocation of the funds.* We distinguish between *non-competitive* allocation (or *block* funding) and *competitive* allocation, e.g., based on (i) submission of research projects/proposals and/or (ii) previous scientific production of candidates. Although there are no general schemes, non-competitive funding is typically provided by the government and intended for public research institutions to cover the salaries of research staff, operating costs and maintenance of infrastructure (e.g., classrooms, laboratories, libraries, etc.) (HFFCE, 2017). It is well known that non-competitive funding is fundamental for the stability and autonomy of entire research institutions (Abramo et al., 2011; Hicks, 2012).
4. *Spending constraints.* For example, time constraints, which impose the use of funds within a certain period, or constraint on the type of expenditure (e.g. for scholarships, support material for teaching, laboratory equipment, consumables, etc.).

The different combination of these four factors may lead to very different funding schemes, each with its own strengths and weaknesses that are often difficult to predict.

In the last three-to-four decades, there has been a tendency in many countries to make public funds increasingly competitive, combined with the gradual diffusion of national exercises for research evaluation (Wang et al., 2018). E.g., in Italy, the public funds that are allocated by the government to public universities include a (non-competitive) fixed share and a (competitive) merit-based share, whose percentage incidence tends to gradually increase over

time: 13% in 2013, 22% in 2017, etc. (Abramo and D'Angelo, 2015; Franceschini and Maisano, 2017).

Due to the economic crisis, Italian universities have suffered a certain decrease in public funding over the last decade, only partly compensated by the increase in private funding (Horta et al., 2008; Abramo et al., 2011; Muscio et al., 2013). Universities and research institutions located in economically/industrially flourishing areas have been able to "exploit" this kind of compensation, certainly more than those in depressed areas.

Returning to the different funding schemes, the scientific literature contains many debates – sometimes very passionate – on their presumed effectiveness (Jacob and Lefgren, 2011; Van Den Besselaar et al., 2017; Wang et al., 2018). Nevertheless, there is often a lack of data on inputs and outputs over a reasonably long period to assess the (un)success of a certain funding scheme. Geuna and Martin (2003) provided a qualitative discussion of the (presumed) advantages and shortcomings of two antithetical research funding schemes: i.e., *competitive* versus *non-competitive*. According to these authors, it would seem desirable – for a generic research institution – to achieve the right mix of competitive and non-competitive funds, with the aim of maximizing the benefits while minimizing the disadvantages. Unfortunately, there are no general rules and the success of a certain research funding scheme is often conditioned by various exogenous factors, such as the socio-economic context, the "health" of the recipient institution, the level of bureaucracy, etc. (Laudel, 2006). The difficulty in studying the (positive and negative) effects of a certain funding policy is also linked to the following limitations:

- The data available are often limited and incomplete;
- The difficulty of defining indicators/proxies that effectively represent important aspects of researchers' activities (e.g., scientific productivity, teaching activity, technology transfer, international scientific reputation, etc.).

It is therefore not surprising to see contradictory opinions on funding policies. For example, Butler (2003) states that the growth of competitive funds encourages productivity, to the detriment of scientific impact. Van Den Besselaar et al., (2017) deny the previous study, showing that the increase in productivity generally "goes hand in hand" with the increase in the scientific impact.

Some authors show that it is more appropriate to allocate funds to "best performers" exclusively and not to a broad base of potential users, in order to maximize results (Auranen and Nieminen, 2010). Other authors show that the excessive concentration of funds can be counterproductive, encouraging the adoption of variable/adaptive methods of selecting the funding recipients, e.g., methods based on an initial recipient selection (with a corresponding preliminary allocation) followed by increasingly stringent recipient selections (and corresponding funding allocations), depending on the results obtained (Reardon, 2007; Fedderke and Goldschmidt, 2015).

Some studies show that – to the right extent – non-competitive funding makes researchers more autonomous and "relaxed", stimulating their creativity and – indirectly – their productivity (Bolli and Somogyi, 2011); in fact, the obsession to seek competitive funds actually absorbs a lot of physical and mental energy. Despite this, there is evidence that the excess of autonomy can result in inefficiency (Wolszczak-Derlacz, 2017).

This paper focuses on an uncommon non-competitive funding initiative involving Politecnico di Torino, hereinafter abbreviated as "PoliTO". PoliTO is an Italian public technical university of good (national and international) reputation, which is composed of about 900 tenured researchers (including assistant/associate/full professors) and provides numerous graduation courses – in Engineering and Architecture mainly – for over 34,000 national/international students (see http://www.polito.it, last accessed on June 2019).

Despite the last-decade period of dire straits of Italian universities, which have experienced a significant decline in terms of public funding (Abramo et al., 2011), the PoliTO cash registers have recently come to have a substantially high liquidity. This liquidity is due to a mixture of causes:

1. In recent years, there has been a massive turnover of research staff, characterized by the large retirement of many senior full professors (with expensive salaries), partly compensated by the recruitment of young researchers (with significantly lower salaries).
2. In the last ten years, the number of students enrolled at PoliTO has been constantly growing, with a consequent increase in revenue, due to tuition fees and the (non-competitive) fixed share provided by the government. On the other hand, the size of the research staff is almost unchanged.
3. In the five-year period 2011-2015, the salaries of all Italian university researchers were temporarily blocked[1], because of a government decree aimed at safeguarding public accounts, trying to contrast the recent international economic crisis. It is roughly estimated that this measure allowed PoliTO to save more than 2M€ per year for five years.
4. Following the recent national research evaluation exercises – i.e., the so-called VQR 2004-2010 and VQR 2011-2014 (Abramo and D'Angelo, 2015; Franceschini and Maisano, 2017) – PoliTO has received a relatively high share of government funding on account of the good results obtained.
5. The university planned to make significant investments in infrastructure for redeveloping some unused public infrastructure (in the city of Turin), with the aim of obtaining new space for teaching activities (e.g., classrooms and laboratories). As a consequence, PoliTO had begun to set aside a certain amount of investment capital. For exogenous reasons, the university then gave up making these investments.

The governmental bodies of PoliTO, with the predominant role of the Board of Administrators[2] (BoA) and the (former) Chancellor, decided in 2016 to use (part of) this liquidity to finance some initiatives for promoting research and teaching activities. A portion was allocated to departments to finance new teaching projects (e.g., improvement of teaching laboratories, acquisition of new hardware/software resources, etc.), while another portion was allocated to departments to finance internal research projects.

In addition, a very large portion of the aforementioned liquidity was allocated directly to individual researchers. Precisely, around the middle of the year 2016, the BoA decided to assign a "lump sum" of 14k€ to each tenured researcher of the PoliTO, to support his/her research activities for the year 2017, without particular constraints, neither on the type of expenditure nor in terms of time.

Because of its universality, this annual funding was denominated "diffused funding"[3] and will henceforth be abbreviated as "DF". The first DF was made accessible since 1st January 2017; the same funding was repeated for each of the following two years, allocating further 14k€ in the year 2018 and 14k€ in the year 2019, for a total of 42k€ to each tenured researcher in three consecutive years.

This substantial non-competitive funding policy is unusual in the current Italian scene and PoliTO researchers are undoubtedly privileged compared to their national counterparts. In fact, we are not aware of any initiatives such as DF in other Italian universities, at least for the moment.

The exceptionality of the initiative is even more evident considering that Italian universities are tendentially centralized and with limited autonomy, which can be extended at the level of

[1] As a rule, salaries are progressively increased depending on the length of service, with two-yearly increments.
[2] In Italian, "Consiglio di Amministrazione".
[3] In Italian, "finanziamento diffuso".

individual researchers (Abramo et al., 2013). In fact, the annual non-competitive share of funding that Italian public universities usually allocate to individual researchers, is around 500-1000 € per year, i.e., one-two orders of magnitude lower than the DF (Muscio et al., 2013)!

The authors believe that the DF plays a role of considerable importance for those dealing with policies for allocating research funds. In fact, this initiative represents an important opportunity to investigate the effects of the substantial allocation of non-competitive funding to individual researchers.

From the above considerations, some interesting research questions arise: "What are the short-term effects of the DF initiative on the scientific output of PoliTO researchers?" and "Are there significant differences between PoliTO researchers and those who have not benefited from the DF?".

The aim of this paper is to compare the scientific output of PoliTO researchers with that of their peers at the national level, trying to provide plausible answers to the previous questions. From a methodological point of view, a sample of PoliTO researchers will be selected, comparing their performance with that of analogous researchers from other Italian universities. The comparison concerns the time window between 2008 and 2018, with special attention to the last two years (i.e., 2017 and 2018) in which PoliTO researchers have benefited from the DF.

The performance will be evaluated considering four different dimensions: (1) "Publishing productivity", (2) "Publishing diffusion/impact", (3) "Journal reputation", and (4) "International research relations". Each of the above dimensions will be represented through appropriate indicators/proxies. The authors are aware that there are other important dimensions that are not strictly focused on scientific publications (e.g., participation in projects, teaching or technology transfer activities), for which the collection of objective and complete data is far more difficult.

The remainder of this paper is organized into three sections. The "Methodology" section describes in detail the data collection/analysis and the indicators used to represent the four analysis dimensions. The "Results" section details the results of the analysis, providing motivated answers to previous research questions. Finally, the "Discussion" section summarizes the original contributions of this research, highlighting possible implications, limitations and ideas for future research.

Methodology

We remind that the objective of this research is comparing the scientific output of PoliTO researchers and that of the researchers from other Italian public universities (hereinafter abbreviated as "Other Universities" or, even more briefly, "OU"), before and after the implementation of the DF. Consistently with this objective, it was decided to make a comparison between groups of researchers in the same disciplines.

In Italy, each tenured researcher from public universities belongs to one-and-only-one specific *discipline*[4], around 370 in all (Abramo et al., 2011; 2013; Franceschini et al., 2013). A complete list accessible at http://cercauniversita.cineca.it/php5/settori/index.php (last accessed on June 2019).

Although the researchers of PoliTO (and other technical universities) are scientifically more homogeneous than those belonging to generalist universities, they can be associated with a number of different disciplines, with significant differences in terms of propensity to publish and cite (Franceschini and Maisano, 2014). For simplicity, among the various PoliTO tenured researchers, we have selected those belonging to the five disciplines reported in Table 1. We

[4] In Italian, "Settore Scientifico Disciplinare", which means "Scientific and Disciplinary Sector".

have limited the selection to researchers that were active in the entire period from 2010 to 2018, thus excluding changes in the staff number (*N*), due to staff retirements, new hires, transfers, etc.. The last two columns of Table 1 report the *N* values of the PoliTO researchers and those of researchers from the OU. Researchers were identified through public directories (MIUR, 2019).

Table 1. Summary of the researchers (from PoliTO and the other universities) considered in the analysis. Researchers are divided into five disciplines.

Discipline	Acronym	Staff number (*N*)		
		Other universities	PoliTO	
A. Design methods for industrial engineering	ING-IND/15	69	4	(5.8%)
B. Manufacturing technology and systems	ING-IND/16	112	16	(14.29%)
C. Industrial mechanical plants	ING-IND/17	120	5	(4.17%)
D. Electrical engineering	ING-IND/31	145	12	(8.3%)
E. Business and management engineering	ING-IND/35	154	6	(3.9%)
	Total	**600**	**43**	**(7.2%)**

Subsequently, the publication output of these researchers was downloaded from the Scopus database. The risk of analysis distortions because of homonymies or other ambiguities is minimized by the database query method, which includes the Scopus IDs of researchers. The documents analysed consist of articles from international journals and conference proceedings indexed by Scopus, which were published in the eleven-year period from 2008 to 2018. For each combination of discipline and year, some indicators – normalized with respect to the staff number – were determined to represent the four dimensions of interest; Table 2 provides a brief description of these indicators.

Table 2. Indicators normalized with respect to the staff numbers (*N*) related to the four dimensions of interest (see Table 1).

Dimension	Indicator	Short description
(1) Publishing productivity	*P*/*N*	The indicator – given by the ratio between the number of journal articles (*P*) by a certain group of researchers and the number of researchers themselves (*N*) – expresses the average level of productivity per capita for a certain year.
(2) Publishing diffusion/impact	*C*/*N*	The indicator – given by the ratio between the total number of citations (*C*) accumulated at the moment of the analysis (i.e., January 2019) by the journal articles of a certain group of researchers and the number of researchers themselves (*N*) – expresses the average level of impact/diffusion per capita for a certain year.
(3) Journal reputation	*Avg. SNIP*	The indicator – which corresponds to the average value of the *SNIP*(*) associated with the journal articles issued in a certain year – is used as a proxy of the reputation of the journals publishing the articles of a group of researchers. As an alternative to *SNIP*, one could use other pre-calculated journal indicators for the Scopus database, such as *CiteScore* (similar to the popular *ISI Impact Factor* of the "competing" Web of Science database) or the *SJR* (Scimago Journal & Country Rank).
(4) International research relations	*F*/*N*	The indicator – given by the ratio between the number of conference articles indexed by Scopus (*F*) by a certain group of researchers and the number of researchers themselves (*N*) – depicts the average presence to conferences per capita, for a certain year. The authors are aware that the conferences indexed by the Scopus database represent a relatively small portion of all the international conferences.

(*) We remark that *SNIP* is an annual field-normalized indicator for ranking scientific journals by Moed (2010a). *SNIP* values related to a particular journal and reference year can be obtained by querying an online application (based on the Scopus database), freely available at http://www.journalindicators.com (last accessed on June 2019). Since the *SNIP* values were not yet available at the time of the analysis for the journals published in 2018, the 2017 *SNIP* values were used.
It is worth remarking that *SNIP* is not available for all the journals indexed by Scopus; articles from journals without *SNIP* were excluded. This operation should not distort the analysis, since the portion of articles excluded is quite stable – usually around 6-8% – for the different groups of papers analysed.

It can be noticed that the indicators representing the dimensions (1), (2) and (4) are normalized by dividing a certain variable (respectively *P*, *C* or *F*) by *N*, in order to allow comparison between groups of researchers of different size.

For each of the five disciplines and each of the above indicators, a pair of time series were determined: the one related to PoliTO researchers and the one related to the researchers of the "OU". Before comparing these pairs of time series, it is necessary to reflect on some "technical" problems:

- The propensity to publish/cite is not necessarily the same for all the disciplines considered (Moed, 2010b; Franceschini and Maisano, 2014).
- Regardless of the discipline, the propensity to publish/cite tends to gradually increase over time. This "inflationary" behaviour is partly related to the increasing pressure to publish and partly related to the constant expansion and growth of the scientific community; for more information, see (Petersen, 2018).
- Older publications have had more time for accumulating citations than the more recent ones.

A further year-by-year normalization of the four indicators in Table 2 was introduced; precisely, each indicator referring to the PoliTO researchers (generically indicated as $x^{(PoliTO)}$) was compared to the corresponding indicator referring to the researchers in the other universities (generically indicated as $x^{(OU)}$), according to the following ratio model:

$$y = \frac{x^{(PoliTO)}}{x^{(OU)}}, \tag{1}$$

Values of y higher/lower than the unit respectively indicate that the performance of PoliTO researchers in the year of interest is higher/lower than that of their counterparts in the rest of the Italian universities. The year-by-year normalization in Eq. 1 was introduced to overcome the aforementioned "technical" problems of comparability among disciplines.

Table 3 summarizes the indicators after this second normalization for each of the four dimensions of interest. As previously described, the normalization is accomplished in two stages: (i) normalization by staff number (if applicable), and then (ii) year-by-year normalization with respect to the OU.

Table 3. Indicators normalized with respect to *N* and then with respect to the OU. These indicators depict the relative performance of groups of Polito researchers with respect to counterparts in the OU.

	Dimension	Indicator	Short description
(1)	Publishing productivity	$\frac{P^{(PoliTO)}/N^{(PoliTO)}}{P^{(OU)}/N^{(OU)}}$	The normalized indicator compares – for a certain discipline and year – the average number of journal articles per capita of PoliTO researchers with that of the counterparts in the OU.
(2)	Publishing diffusion/impact	$\frac{C^{(PoliTO)}/N^{(PoliTO)}}{C^{(OU)}/N^{(OU)}}$	The normalized indicator compares – for a certain discipline and year – the average number of citations per capita accumulated by the journal articles of PoliTO researchers with that of the counterparts in the OU.
(3)	Journal reputation	$\frac{Avg.SNIP^{(PoliTO)}}{Avg.SNIP^{(OU)}}$	The standardised indicator compares – for a certain discipline and year – the *Avg. SNIP* value of the journal articles of PoliTO researchers with that of the counterparts in the OU.
(4)	International research relations	$\frac{F^{(PoliTO)}/N^{(PoliTO)}}{F^{(OU)}/N^{(OU)}}$	The normalized indicator compares – for a certain discipline and year – the average number of conference articles per capita of PoliTO researchers with that of the counterparts in the OU.

The choice to consider the indicators in the eleven-year period from 2008 to 2018 is motivated by the following reasons:

- Since the indicators in the first nine years (2008-2016) are not influenced by the DF, they will allow to estimate the "natural performance"[5] of PoliTO researchers with respect to their counterparts in the OU.

[5] The "natural performance" may be defined as the performance that PoliTO researchers tended to exhibit before benefiting from the DF.

- This estimate will constitute a sort of reference for assessing whether in the last two years (i.e., 2017 and 2018, which are potentially influenced by the DF) PoliTO researchers have been performing differently than in the nine previous years.

The authors are aware that two years is a relatively limited period to evaluate possible effects of the DF, considering the production time of scientific publications and the maturation time of the respective citations (Van Den Besselaar et al., 2017). Nevertheless,

- Engineering is a field where the production of scientific publications is faster than other fields (Moed, 2010b);
- Although the DF has been disbursed since 2017, the initiative was officially approved around mid-2016. It therefore seems reasonable to assume that this information may have influenced the behaviour/attitude of PoliTO researchers even before 2017, when they actually had access to the funding capital.

Results

The Scopus database was queried, extracting the data concerning the groups of researchers (of PoliTO and the OU) and constructing the indicators introduced in the section "Methodology", for each of the four dimensions of interest.

Let us start by considering the dimension (1) "Publishing productivity". For the four disciplines, the productivity of PoliTO researchers is generally in line with that of the counterparts in the OU. As an example, consider the graph in Figure 1, about the discipline "B. Manufacturing technology and systems". It can be noted that the PoliTO profile is a little more nervous than the corresponding OU profile; this is not surprising, considering the significantly lower number of PoliTO researchers, which makes the annual estimates of their productivity per capita more dispersed than that of the corresponding researchers in the OU.

From Figure 1(1), we can also appreciate a certain inflationary phenomenon, which leads to a certain growth of the publishing activity over time, in line with what is documented in (Petersen, 2018). In addition, we observe systematic differences between researchers belonging to different disciplines: for example, PoliTO researchers in the discipline "D. Electrical Engineering" seem to perform significantly better than their counterparts in the OU, while PoliTO researchers in the discipline "C. Industrial mechanical plants" tend to perform worse.

The considerations seen for the dimension (1) "Publishing productivity" can be extended to the other three dimensions, i.e., (2) "Publishing diffusion/impact", (3) "Journal reputation", and (4) "International research relations". Again, the initial data were normalized in two stages: (i) based on the staff number (*N*) and (ii) with respect to the OU. The only exception is the indicator "*Avg. SNIP*", which represents the dimension (3) "Journal reputation": being this indicator *size-independent*, the first normalization stage is not necessary (Franceschini et al., 2013).

Inflationary phenomena can also be observed for dimensions (3) and (4)[6], leading to a gradual increase in productivity and citation impact. See for example the graphs in Figures 1(3) and 1(4), which show the curves related to the above dimensions for discipline "B. Manufacturing technology and systems". Similar results can be observed for the other disciplines. The year-by-year normalization with respect to the OU makes it possible to "purge" the aforementioned inflationary phenomena, obtaining deflated data (see Figure 2(a)).

For each analysis dimension, it is finally possible to aggregate the time series relating to the five individual disciplines into a single (aggregated) time series, depicting the overall performance of PoliTO researchers from the perspective of the dimension of interest. The

[6] As for dimension (2), this phenomenon is partly hidden since the so-called "citation inflation" is compensated by the lower time available for citation accumulation of the more recent journal articles.

aggregation can be made through their weighed sum, using as weight the staff number of the PoliTO researchers for a certain d-th discipline ($N_d^{(PoliTO)}$):

$$w = \frac{\sum_{\forall d} y_d \cdot N_d^{(PoliTO)}}{\sum_{\forall d} N_d^{(PoliTO)}}, \tag{2}$$

w being the aggregate value for a certain year and the "d" subscript denoting a certain discipline ($d \in \{\mathrm{A}, \mathrm{B}, \mathrm{C}, \mathrm{D}\}$, as illustrated in Table 1).

The results of the aggregated indicators for the four dimensions of interest are represented in Figure 2(b) and will be further recalled in Table 4.

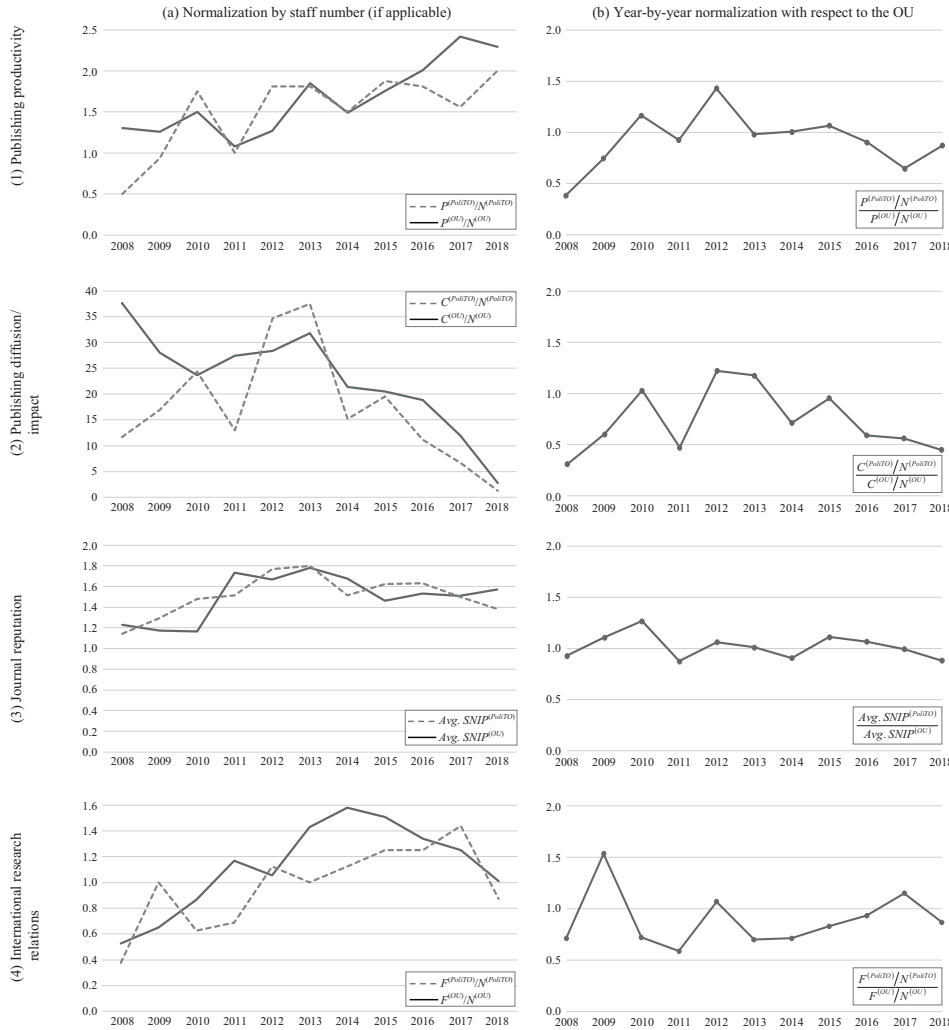


Figure 1. Graphic representation of the (normalized) data, with reference to the discipline "B. Manufacturing technology and systems".

Focussing on the corresponding aggregated-indicator charts in Figure 2(b), it can be noticed that – being purged of the above-described inflationary effects too – they show a (presumably) random trend. To confirm this, an Anderson Darling normality test at 95% for each of the four dimensions was carried out. In all cases, the null hypothesis that the aggregated indicator is normally distributed could not be rejected. This test was performed using the data between 2008 and 2016, since they are not influenced by the DF and therefore depict the "natural" overall performance of PoliTO researchers.
Focussing on the results related to the years 2017 and 2018, which are potentially influenced by the DF, it would seem that there is no apparent increase in the performance of PoliTO researchers compared to their counterparts in the OU for all the dimensions of interest. In several cases, it actually seems that there is a decreasing trend!

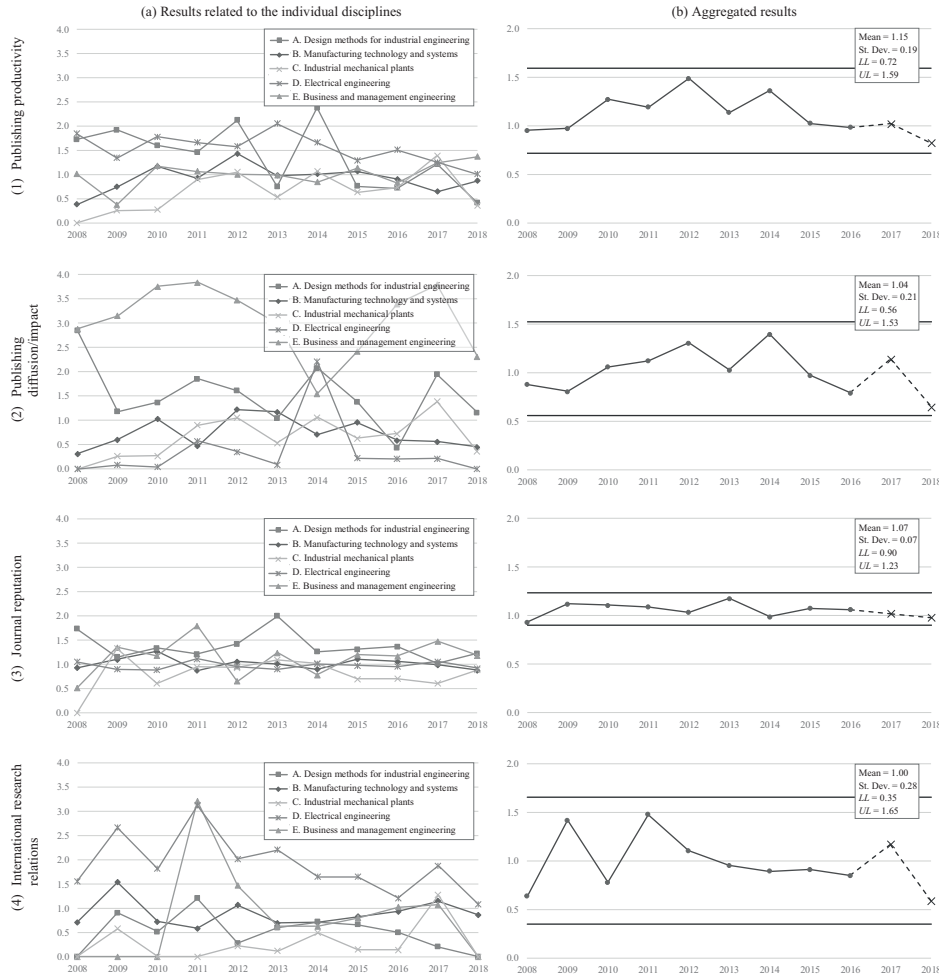


Figure 2. Charts related to the (normalized) time series related to the four analysis dimensions.

The previous trends can be examined more rigorously by means of a statistically sound test. For each of the four time series, it is constructed a 95% confidence interval, which should include the realizations of a random variable with $w \sim N(\mu, \sigma^2)$, μ and σ^2 being both unknown.

To estimate the "natural performance" of the PoliTO researchers, we only used the data in the nine-year period 2008-2016, since they are certainly not influenced by the DF.
Compatibly with the available data, the best possible estimator of μ is:

$$\hat{\mu} = \overline{w} = \frac{1}{9} \cdot \sum_{i=2008}^{2016} w_i \ . \tag{3}$$

Since the sample of data is small (in fact, it includes only nine observations), the sample variance

$$s^2 = \frac{1}{9-1} \cdot \sum_{i=2008}^{2016} \left(w_i - \overline{w}\right)^2 , \tag{4}$$

systematically underestimates σ^2 (Ross, 2009). The $\overline{w}$ and s values related to the four aggregated time series are reported in Table 4 (see the columns "Mean" and "St. Dev." respectively). A 95% confidence interval – delimited respectively by a lower limit, (*LL*) and an upper one (*UL*) – can be constructed using the Student *t* distribution:

$$\begin{aligned} UL &= \overline{w} + t_{\nu, 1-\frac{\alpha}{2}} \cdot s \\ LL &= \overline{w} - t_{\nu, 1-\frac{\alpha}{2}} \cdot s \end{aligned} , \tag{5}$$

being $\nu = 9–1 = 8$, $\alpha = 5\%$ and therefore $t_{\nu, 1-\frac{\alpha}{2}} = t_{\nu,\, 97.5\%} = 2.306$.

For each of the analysis dimensions, the resulting values of *LL* and *UL* are reported in the last two columns of Table 4 and graphically represented in the graphs in Figure 2(b).
It can be noticed that all the time series are contained in these limits in the period 2008-2016, reinforcing the hypothesis that they do not show any non-random trend. More interestingly, even the data for the years 2017 and 2018 are all within the relevant confidence-interval limits. This means that there are no statistically significant differences between the results of the PoliTO researchers in the last two years (2017 and 2018) with respect to the previous nine years (2008 to 2016), for each of the analysis dimensions. In other words, it seems that there are no significant effects of the DF on the dimensions of interest.

Table 4. Aggregated time series related to the four dimensions of analysis and relevant statistics.

Dimension	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	Mean†	St. Dev.†	*LL*	*UL*
(1) Publishing productivity	0.95	0.97	1.27	1.19	1.48	1.14	1.36	1.03	0.98	1.02	0.83	1.15	0.19	0.72	1.59
(2) Publishing diffusion/impact	0.88	0.81	1.06	1.13	1.31	1.03	1.40	0.98	0.79	1.14	0.65	1.04	0.21	0.56	1.53
(3) Journal reputation	0.93	1.12	1.11	1.09	1.04	1.18	0.99	1.08	1.06	1.02	0.96	1.07	0.07	0.90	1.23
(4) International research relations	0.63	1.42	0.78	1.47	1.10	0.95	0.89	0.91	0.85	1.17	0.59	1.00	0.28	0.35	1.65

(†) These statistics have been calculated taking into account only the data in the period not affected by the DF (i.e., from 2008 to 2016).

Discussion

The analysis showed that the DF did not produced any significant increase in any of the four dimensions of interest. However, this result should be considered with prudence, in view of some significant limitations of this research:

- The analysis was based on a relatively limited sample of PoliTO researchers, both in terms of size (43 out of around 900 tenured researchers) and discipline (5 out of around 40 disciplines).
- Two years (i.e., 2017 and 2018) is a relatively limited period to evaluate the (missing) effects of the DF, especially considering the intrinsic inertia associated with the production of scientific publications and the citation accumulation. However, the fact remains that in the Engineering field, the production of publications is relatively faster than in other fields,

which makes the results of the analysis interesting, at least at a preliminary level (Moed, 2010b).

- The analysis dimensions are strictly focused on scientific publications, which implies neglecting other important scientific outputs for the life of a researcher, such as participation in (national/international) projects, teaching, or technology transfer activities. The reason for this "under-representation" is the difficulty of collecting objective and complete data on these other scientific outputs.

Returning to the DF, the author believe that it certainly gave the PoliTO researchers a certain sense of peace of mind, eliminating the hassle of increasing the indicators of research output, at least in the short term (Laudel, 2006). This is due not only to the amount of the funding but also to the rules that govern it:

- There are no time restrictions on the use of funds;
- There are no restrictions on the type of expenditure (e.g., distinguishing between "noble" expenditure, such as acquisition of research equipment or payment for research missions, and less noble expenditure, such as purchase of consumables, computer equipment, office furniture, etc.).
- There is no obligation for the researcher to co-finance the expenditure, which would be an important deterrent for the purchase of material that is not strictly important for research.
- There are no minimum results for the receipt of new annual funds (i.e., 14k€ per year), depending on the research activity resulting from the use of the funds received in previous years.

The introduction of some constraints to these relatively "libertarian" rules could probably lead to more concrete results (Bolli and Somogyi, 2011). Nevertheless, it is not excluded that the initiative may be effective in the medium to long term, as it is. In this respect, we plan to investigate whether the results of this preliminary short-term study will be confirmed in the medium-long term, perhaps considering a larger sample of researchers from PoliTO and the OU.

References

Abramo, G., Cicero, T., D'Angelo, C.A. (2011). The dangers of performance-based research funding in non-competitive higher education systems. *Scientometrics*, 87(3), 641-654.

Abramo, G., Cicero, T., D'Angelo, C.A. (2013). The impact of unproductive and top researchers on overall university research performance. *Journal of Informetrics*, 7(1), 166-175.

Abramo, G., D'Angelo, C.A. (2015). The VQR, Italy's second national research assessment: Methodological failures and ranking distortions. *Journal of the Association for Information Science and Technology*, 66(11), 2202-2214.

Auranen, O., Nieminen, M. (2010). University research funding and publication performance—An international comparison. *Research Policy*, 39(6), 822-834.

Bolli, T., Somogyi, F. (2011). Do competitively acquired funds induce universities to increase productivity?. *Research Policy*, 40(1), 136-147.

Butler, L. (2003). Explaining Australia's increased share of ISI publications—the effects of a funding formula based on publication counts. *Research policy*, 32(1), 143-155.

Fedderke, J.W., Goldschmidt, M. (2015). Does massive funding support of researchers work?: Evaluating the impact of the South African research chair funding initiative. *Research Policy*, 44(2), 467-482.

Franceschini, F., Maisano, D., Mastrogiacomo, L. (2013). Evaluating research institutions: the potential of the success-index. *Scientometrics*, 96(1), 85-101.

Franceschini, F., Maisano, D. (2014). Sub-field normalization of the IEEE scientific journals based on their connection with Technical Societies. *Journal of Informetrics*, 8(3), 508-533.

Franceschini, F., Maisano, D. (2017). Critical remarks on the Italian research assessment exercise VQR 2011–2014. *Journal of Informetrics*, 11(2), 337-357.

Geuna, A., Martin, B.R. (2003). University research evaluation and funding: An international comparison. *Minerva*, 41(4), 277-304.
HEFCE (Higher Education Funding Council for England) (2017). *Guide to funding 2017-18: How HEFCE allocates its funds*, http://dera.ioe.ac.uk/29341/ [retrieved on January 2019].
Hicks, D. (2012). Performance-based university research funding systems. *Research policy*, 41(2), 251-261.
Horta, H., Huisman, J., Heitor, M. (2008). Does competitive research funding encourage diversity in higher education?. *Science and public policy*, 35(3), 146-158.
Jacob, B.A., Lefgren, L. (2011). The impact of research grant funding on scientific productivity. *Journal of public economics*, 95(9-10), 1168-1177.
Laudel, G. (2006). The art of getting funded: how scientists adapt to their funding conditions. *Science and Public Policy*, 33(7), 489-504.
MIUR (Ministero dell'Istruzione dell'Universita` e della Ricerca) (2012). MIUR Settori scientifico-disciplinari, http://cercauniversita.cineca.it/php5/settori/index.php [retrieved on January 2019].
Moed, H.F. (2010a). CWTS crown indicator measures citation impact of a research group's publication oeuvre. *Journal of Informetrics*, 3(3), 436-438.
Moed, H.F. (2010b). Measuring contextual citation impact of scientific journals. *Journal of informetrics*, 4(3), 265-277.
Muscio, A., Quaglione, D., Vallanti, G. (2013). Does government funding complement or substitute private research funding to universities?. *Research Policy*, 42(1), 63-75.
Petersen, A.M. (2018). Multiscale impact of researcher mobility 15 *Journal of The Royal Society Interface*, doi: 10.1098/rsif.2018.0580.
Reardon, S. (2007). NIH to limit the amount of grant money a scientist can receive. Nature News (3 May 2017), doi: 10.1038/nature.2017.21930.
Ross, S.M. (2009). *Introduction to probability and statistics for engineers and scientists*. Academic Press.
Van Den Besselaar, P., Heyman, U., Sandström, U. (2017). Perverse effects of output-based research funding? Butler's Australian case revisited. *Journal of Informetrics*, 11(3), 905-918.
Wang, J., Lee, Y.N., Walsh, J.P. (2018). Funding model and creativity in science: Competitive versus block funding and status contingency effects. *Research Policy*, 47(6), 1070-1083.
Wolszczak-Derlacz, J. (2017). An evaluation and explanation of (in) efficiency in higher education institutions in Europe and the US with the application of two-stage semi-parametric DEA. *Research Policy*, 46(9), 1595-1605.

Identification of the research on warfare and health, 1946-2017

Grant Lewison[1],Marian Abouzeid[2,3,4], Ammar Sabouni[3], Manal El Zalabany[3], Samer Jabbour[3,4] and Richard Sullivan[1]

[1] *grantlewison@aol.co.uk; richard.sullivan@kcl.ac.uk*
School of Life Sciences and Medicine, King's College London, London SE1 9RT (United Kingdom)

[2,3,4] *marian@cfhl.org.au*
[2]Centre for Humanitarian Leadership, A Deakin University – Save the Children Australia partnership, Burwood 3220, VIC (Australia)

[3]The Lancet-AUB Commission on Syria (Lebanon)
[4]Faculty of Health Sciences, American University of Beirut (Lebanon)
[3]*lcs-elzalabany@aub.edu.lb, lcs-sabouni@aub.edu.lb*
[3,4] *sj22@aub.edu.lb*
[4]Faculty of Health Sciences, American University of Beirut (Lebanon)

Abstract
In order to characterise the scope, thematic content and evolution of the field of literature on health and armed conflict, we undertook a preliminary study to identify the body of research undertaken since the Second World War. This involved the iterative development of a complex filter which was applied to the Web of Science. We included papers on the effects of war on both soldiers and on civilian populations, both in the combat zone and when displaced. In total, we found 10,467 papers, after manual removal of many irrelevant ones. The large majority came from the USA, but several countries were much more active in research on warfare and health than in biomedical research overall, notably the small countries that had suffered most from recent warfare. Because of the much-increased volume of biomedical research since year 2000, there were many papers in the last two decades, especially on mental disorders rather than physical wounds, and their effects on soldiers rather than civilians and refugees. Japanese papers were the most highly cited; many were on the effects of ionising radiation.

Conference Topic
Scientific-scholarly internationalisation, collaboration and mobility: relationship to disease burden

Introduction
This study is a collaborative initiative between the Research for Health in Conflict - Middle East and North Africa programme (R4HC-MENA) and the Lancet - American University of Beirut Commission on Syria (LCS). Briefly, R4HC was set up by the UK's Economic and Social Research Council to run from October 2017 to December 2021 (https://r4hc-mena.org/). R4HC-MENA's objective is to build research and policy capacity in conflict-affected areas, focusing on health, the political economy of health, and complex non-communicable diseases such as mental health and cancer, and facilitate more effective translation of research into policy. The intention is to create a new, networked community with practice, policy and health systems financing and planning informed by the best available evidence. The main focus of the programme will be in Jordan, the Occupied Palestinian Territories, Lebanon and Turkey, as well as conflict-affected populations in Syria, Iraq, Libya and Yemen. The LCS was established in 2016 in response to the profound loss of life, human suffering, destruction of health and social systems and mass population displacement in Syria, and aims to produce the best possible science in relation to health and the Syrian conflict, and to mobilise global action to strengthen the response to the conflict, with a particular focus on health (https://website.aub.edu.lb/lcs/Pages/home.aspx). The first Lancet Commission to be headquartered and led by an institution in the Global South, the LCS is a highly collaborative research and policy translation initiative that examines a full spectrum of issues related to health and war. Initially focussed on the Syrian conflict and its local, regional and global effects, many of the issues arising and questions warranting

consideration have led to a broadening of focus from Syria to other wars and armed conflicts in the region and beyond.

We conducted a bibliometric study of research conducted world-wide during the period 1946 to 2017, in order to identify the body of literature on health and armed conflict, gauge thematic evolution of this field of research over time, assess issues of localisation of research, and to gauge information gaps warranting further attention. This paper describes the methodology used to characterise a field of research that includes contributions from numerous diverse disciplines, and outlines the assumptions made, boundaries drawn and sub-filters developed to identify this body of research and its components. Some preliminary results are also given, but the main results and consequent conclusions will be given later.

Methodology and Analysis

The first task was to define the scope of the study. We chose to examine research both on the effects of warfare on the health of the military personnel taking part and on the civilian population in the theatres of war, but not papers on the problems faced by the families of the deployed soldiers, as we regarded them as being too indirectly connected to warfare. Warfare creates many refugees, and we included papers that described their health problems, but excluded those that arose when they were housed in refugee camps as they often have causes other than the armed conflict from which they have fled.

Not all the health effects on soldiers are a direct result of offensive action: until the Second World War more fighting men died or become ill as a result of disease and difficult living conditions (Cirillo, 2008). The same applies to the civilian population who may suffer from a severe lack of affordable food and drinking water, from exposure, and from restricted access to medical treatment, because of logistical problems and destruction of infrastructure caused by conflict.

We also considered the health problems faced by prisoners of war. Some casualties result from the use of landmines and other unexploded ordnance, which are a consequence of military activity. We included all these situations, but were careful to exclude the *sequelae* of natural disasters (floods, fires, earthquakes, and droughts), and also of low-level violence including gang violence, domestic aggression, and casualties attributable to acts of terror, such as those perpetrated by suicide bombers.

Papers (articles, notes and reviews) in the Web of Science (WoS, © Clarivate Analytics) were identified by means of a complex filter that needed seven iterations by AS and GL in order to achieve a good level of both precision and recall (specificity and sensitivity). The main positive filter consisted of two sets of title words. One set derived from military activities, such as:

> *armed-conflict, battlefield, combat-related, deploy*, Enduring Freedom* and the Journal: *Conflict and Health* (set 1).

A second parallel set was based on geographical combat zones, such as:

> *Afghanistan, Cambodia, Darfur, East-Timor, Gulf*

plus a restricted set of military and personnel words, such as:

> *army, bomb, combat, deploy*, refugee, soldier, veteran.* (set 2)

These were combined with a third, health set, with words such as:

> *anxiety, casualt*, death, fever, hygien*, injur*, medical, nursing* (set 3)

but subject to a fourth set of "no" title words that would remove irrelevant papers, such as:

> *airbag, breast, cigarette, dengue, earthquake, fertility, game* (set 4)

which had been found on an individual basis as the filter was being developed. The final filter then consisted of ((set 1 or set 2) + set 3) - set 4. However, we found that the word "deploy" could be used in several different senses, in particular for the use of medical devices and

techniques, so these papers also needed to be manually removed from the database. We also needed to remove papers with "liberat*", "war on/against cancer", and "soldier fly" from the database, another 158 in total. The final value of the precision (p, specificity), based on markings of 5 x 200 paper samples was **0.897 ± 0.09**. We attempted to determine the recall (sensitivity) by means of the rule-of-three method (Lewison, 1996), but it proved difficult to obtain an appropriately random sample of the papers from eponymously-named departments (a combination of both military and medical address terms). It appeared that almost all the papers from these departments were irrelevant to the subject, so the conclusion was that the recall was quite high, but that it could not be estimated satisfactorily.

Subsequently, as it appeared that there were relatively few papers with "Syria" or "Syrian" in the title, despite the profound health consequences of the civil war in that country, we examined all the papers in the WoS with these words in their titles in the five years from 2013 to 2017. Of the 1391 such papers, only 179 were clearly relevant to warfare and health, and 93 had been identified by the filter. Examination of the 86 non-identified papers suggested a number of additional search terms, and the revised (seventh) version of the filter captured 57 of them. This means that only 29 of the 179 relevant papers were missed, or 16% For these five years, the final version of the filter identified a total of 2207 papers, but 130 were found on inspection to be irrelevant, most of them because the word "deploy*" was used for other subject areas. So the final total was **10,467** papers, and from analysis of the "Syrian" papers, the recall (r, sensitivity) could be calculated as 150/179 = **0.838**. The calibration factor, p/r = 1.07, so the true total (if the filter had been perfect) would have been 11,204 papers in the WoS.

Although it might be thought better to identify relevant papers by means of words or phrases also in the abstracts, or author keywords, this is likely to capture few additional papers, and it significantly degrades the filter's precision (Lewison, 2011). In cancer research, which is rather easier to define than warfare and health, the use of abstracts and keywords added only 4% to the recall, but the precision was degraded from 0.93 to 0.57. It appears that titles alone are sufficiently descriptive of the contents of the paper.

Allocation of papers to countries

Once the final set of papers had been created in the form of an Excel spreadsheet, it was analysed in several different ways. The first procedure was to parse the address field, and mark each paper with the fractional count of countries that were present. Because the time period being examined was so long (papers published between 1946-2017), some countries had changed their names, notably several of those in eastern Europe and parts of the former Soviet empire, including East Germany. This was of particular importance for the former Yugoslavia, as there were many papers from Croatia and other component countries as a result of the wars of the early 1990s. Papers from before these countries became independent were classified with reference to the major cities in the respective countries, for example Pristina, Prizren, Mitrovica and Gjakove in Kosovo. Some papers had no addresses: from 1946 to 1955 more than 90%, and all the papers from 1956-65, did not have them. The percentage of papers from 1966-72 without addresses was above 80%, but in 1973 and later years, almost all the papers had addresses. [This appeared to be a feature of the biomedical papers in the WoS, and not just the ones on warfare and health.] Consequently, most of our analysis has been confined to the papers from 1973-2017, grouped in nine five-year periods.

We wished to "normalise" the outputs of warfare and health (WARHE) papers by comparing them with all biomedical research (BIOMED) in these years. Biomedical papers were identified by means of an address-based filter (Lewison & Paraje, 2004). They were far too numerous (over 16 million) for us to be able to download their details, so comparisons of the percentage presence of different countries in the two sets of papers were made on the basis of

integer (or whole) counts of countries present in the addresses. A few countries, badly affected by war, had a much higher presence in the WARHE set than in the BIOMED set: we limited the comparison to those with at least three WARHE papers on a fractional count basis.

International collaboration can be measured in several different ways (Savanur & Srikanth, 2006; Liao &Yen, 2010; Dozier *et al*., 2014; Kim *et al*., 2014). The simplest method is to determine the difference between the sum of the individual country contributions and the total number of papers. For the WARHE set, we were also able to see which pairs of countries collaborated most closely in the period since 1973. This was done on the basis of whether country A co-authored papers with country B more than would have been expected if it had chosen its international partners purely on the basis of country B's presence in the field *minus* country A's contribution. [For example, Canada published 269 papers with a domestic contribution of 193.3 papers and a foreign one of 75.7 papers, and the US contribution to these was 48.2 papers, or 63.7% of them. The US total output was 4934 papers out of a non-Canadian total of (9590 - 193.3) = 9396.7, or 52.5%. So it was one of Canada's preferred partners by a factor of 63.7 / 52.5 = 1.21. Other countries, of course, would be less preferred.]

Analysis of the title words

Of particular interest was to see which papers described research on individual wars, and to determine the time-scale of the papers that followed them. This was conveniently done by means of a search for country names in the titles of the papers, or sometimes for other words. For example, there were many papers analysing the effects of ionising radiation (previously called "atomic") on survivors from the nuclear bombs dropped on Japan in 1945 that ended the Second World War. The other major conflicts on which there were many papers were the Korean War (1950-53), the Vietnam Wars (1946-54 against France, and 1965-75 against the USA and others), the Yom Kippur war between Israel and its neighbours (October 1973), those in Bosnia (1992-95), Croatia (1991-95), and Kosovo (1998-99), the conflicts between Iran and Iraq (1980-88), the two USA-led invasions of Iraq (1991 and 2003), and the continuing war in Afghanistan (from 2001 to the present). There were also wars in Africa, some of them long-lasting civil wars, such as the ones in Liberia (1989-96), Sierra Leone (1991-2002), Uganda (mainly 1986-2006), and Rwanda (1990-94), and the wars of independence in Angola (1961-74) and Mozambique (1964-75), and in East Timor (1975-99). The numbers of papers in each five-year period since 1973 were tabulated, and then divided by the total in order to give a percentage of papers in each quinquennium.

Analysis of the words in the titles also allowed us to see how the focus of research shifted over the 45-year period from 1973 to 2017, from mainly physical wounds to post-traumatic stress disorder (PTSD) and other mental disorders such as depression, anxiety and suicide. We also looked at whether the research covered the health effects of war on refugees and other displaced persons (both immediate effects and those occurring years later when they had moved to safe countries), as compared with the effects on soldiers, recruits and veterans, and on medical and nursing personnel. The relevant papers were also identified by means of title words.

Finally, we determined the citation counts to papers over a five-year period beginning with the publication year (actual citation impact, ACI). This is a compromise time window between the need for immediacy and for the peak of citation counts to have occurred. We calculated the mean ACI value for 13 leading countries, and also the numbers of papers in the top 5% of ACI values overall (31 cites; actually 5.17% of the 7852 citable papers) from each country as a percentage of its citable papers. This, when multiplied by 100, gives its "worldscale - WS" value, relative to the norm of 100 (Lewison *et al*., 2007).

Preliminary results

Outputs of papers, time-line and relation to biomedical research

The final data set comprised 10,467 papers. The WoS did not cover addresses fully for biomedical papers prior to 1973, so the comparison of the WARHE papers with the BIOMED set, shown in Figure 1, only starts in 1973 and is shown as five-year totals from 1973 to 2017. Two conclusions are clear. Research on WARHE jumped after 2007, and stayed high; and WARHE as a percentage of biomedical research increased from 0.032% in 1973-77 to 0.080% in the last quinquennium.

Table 1 shows the ratio (if it exceeded 0.5) of the percentage presences of the individual countries with at least three WARHE papers in this set to those in the biomedical set in the years 1973-2017. This shows which countries were doing the most warfare and health research during this period, relative to their biomedical research output. The 11 leading countries in the first column are clearly ones that have been heavily involved in wars, or have many war-related refugees.

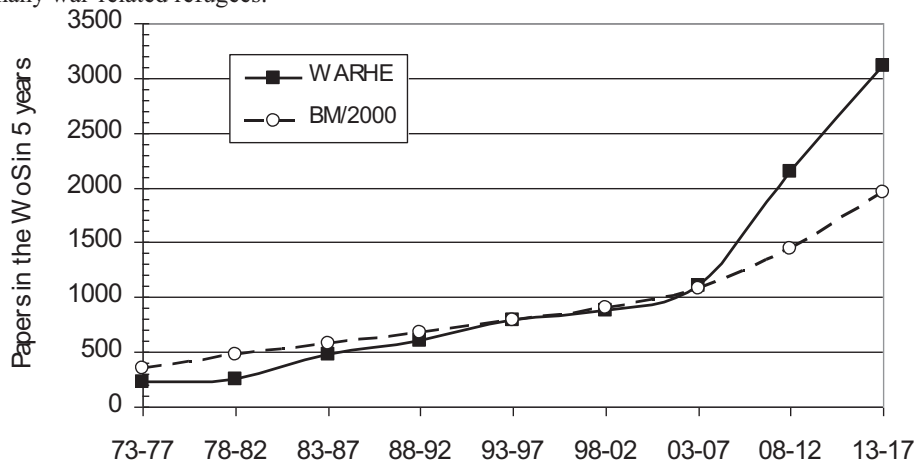


Figure 1. Growth of outputs of papers on warfare and health (WARHE) and biomedical research (divided by 2000) from 1973 to 2017.

Table 1. Ratio of percentages of countries in warfare and health papers to their percentages of biomedical research papers in the years 1973 to 2017.

Country	*Ratio*	*Country*	*Ratio*	*Country*	*Ratio*	*Country*	*Ratio*
Afghanistan	65.7	Libya	9.71	Indonesia	1.76	Nigeria	0.91
Palestine	49.6	Sudan	8.11	Kenya	1.73	Turkey	0.77
Liberia	45.8	Sri Lanka	5.90	USA	1.56	New Zealand	0.76
Bosnia & H.	42.0	Jordan	5.39	Iran	1.42	Netherlands	0.73
Kosovo	30.6	Colombia	4.27	Saudi Arabia	1.23	Thailand	0.72
Iraq	23.4	Israel	4.25	Australia	1.13	Switzerland	0.62
Croatia	17.2	Zimbabwe	4.18	Norway	1.12	Denmark	0.62
Lebanon	16.1	Ethiopia	4.01	South Africa	1.11	Canada	0.61
Uganda	15.4	Kuwait	4.00	UK	1.05	Germany	0.59
Rwanda	13.0	Serbia	3.74	Egypt	0.98	Finland	0.57
Syria	10.0	Vietnam	1.93	Pakistan	0.95	France	0.51

International collaboration

For biomedical research in 1973-2017, the number of papers was 16,498,857 and the sum of the individual country totals was 20,454,550. The ratio between these two numbers is 1.24.

The corresponding totals for the WARHE papers were 9206 papers and the sum of country integer counts was 11,011. This gives a ratio of 1.196, which is smaller, meaning that there is less international collaboration than in biomedical research. For WARHE literature, this ratio increased from 1.04 in 1973-92 to 1.30 in 2013-17. [The differences are statistically highly significant, p << 0.1%, on the Poisson distribution with one degree of freedom.] The amount of international collaboration has increased quite rapidly over the last 45 years, see Figure 2.

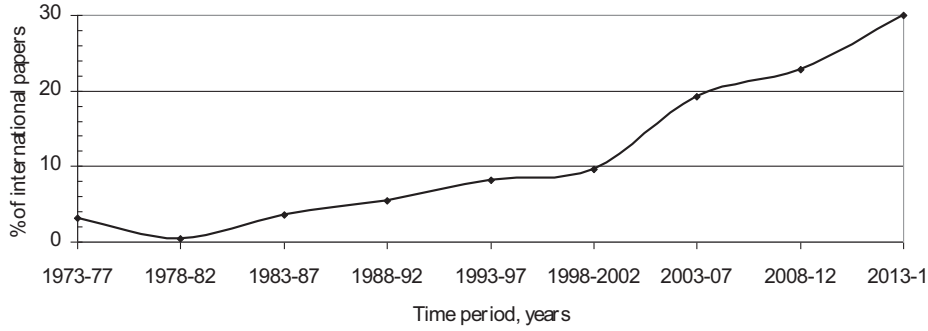


Figure 2. Percentages of WARHE papers with more than one country named in the address field, 1973-2017.

Collaboration between the leading 12 countries was determined to see how much took place, and whether the numbers of joint papers were statistically more, or less, than the numbers expected. The matrix in Table 2 is not symmetrical. The table shows the results for each partner country (in the columns) for a given country (in the rows). Thus for the USA, the Canadian presence of 31 papers is highly significant (p < 1%), but for Canada the US presence of 48 papers is not statistically significant.

Table 2. Numbers of collaborative papers between two of the 12 leading countries in WARHE, 1973-2007, fractional counts. Results are not significantly different from the expected values except those shaded and in small type (smaller) or in large type (larger). Values in bold are statistically significant at p < 1%; those in normal type are significant at p < 5%.

Country	*ISO*	*US*	*UK*	*IL*	*DE*	*AU*	*HR*	*FR*	*CA*	*JP*	*NL*	*TR*	*IR*
USA	*US*		49	27	36	14	**6.1**	**8.1**	**31**	17.2	**22**	2.5	3.8
UK	*UK*	**65**		3.2	9.1	9.4	8.4	6.1	5.0	1.5	**9.1**	1.3	0.5
Israel	*IL*	29	3.3		1.6	0.2	1.7	1.7	2.1	0.0	0.3	0.3	0.0
Germany	*DE*	63	10	1.4		2.2	4.6	2.4	0.6	0.7	4.4	0.1	0.5
Australia	*AU*	20	8.0	0.1	2.0		0.0	1.0	1.8	0.5	1.6	0.5	2.5
Croatia	*HR*	**3.4**	6.0	1.0	3.5	0.0		0.0	0.0	0.0	2.8	0.2	0.0
France	*FR*	**9.6**	5.0	1.3	0.9	0.3	0.0		0.3	0.0	0.2	0.0	0.5
Canada	*CA*	48	6.7	2.6	1.1	1.0	0.0	0.3		0.0	1.6	0.0	0.6
Japan	*JP*	16	1.5	0.0	1.3	0.5	0.0	0.0	0.0		0.0	0.0	0.0
Neth'ds	*NL*	**22**	10	0.3	4.6	0.8	2.7	0.5	1.9	0.0		2.0	0.0
Turkey	*TR*	2.3	1.2	0.7	0.1	0.3	0.6	0.0	0.0	0.0	1.4		0.0
Iran	*IR*	2.4	0.5	0.0	0.5	1.5	0.0	0.5	0.2	0.0	0.0	0.0	

The changing focus of the papers with time

The titles of the papers often mentioned the theatre of war that the research covered, and the names of the countries (occasionally in adjectival form) in the nine quinquennia from 1973-77 are shown in Figures 3-7 as percentages of the total numbers of WARHE papers in each period. Note that the scale of the five charts differs: Figures 4 and 8 have 25% as maxima, showing how the *oeuvre* is dominated by US papers on the effects of its major wars in Viet Nam, Iraq and the continuing conflict in Afghanistan. Many of the papers on the latter two conflicts cover the effects on the army in either or both. It was difficult to distinguish between papers on the two US-led Iraq invasions, so there may have been some double counting. In Figure 6, the first set of papers refers to the war between Iran and Iraq in 1980-88.

Figure 8 shows the distribution of papers on particular subject areas, again taken from the words in their titles: radiation (atomic, nuclear); wounds (injury); anxiety (depression, suicid*); and PTSD (post-traumatic).

Next, we show in Figure 9 the effects of warfare on the different classes of personnel involved: soldiers (veterans, conscripts, recruits, reservists); prisoners (captiv*); civilians (children, elderly, victims, persons); refugees (displaced persons); and medics (doctors, surgeons, nurses, physicians, psychologists, psychiatrists, dentists, first attenders).

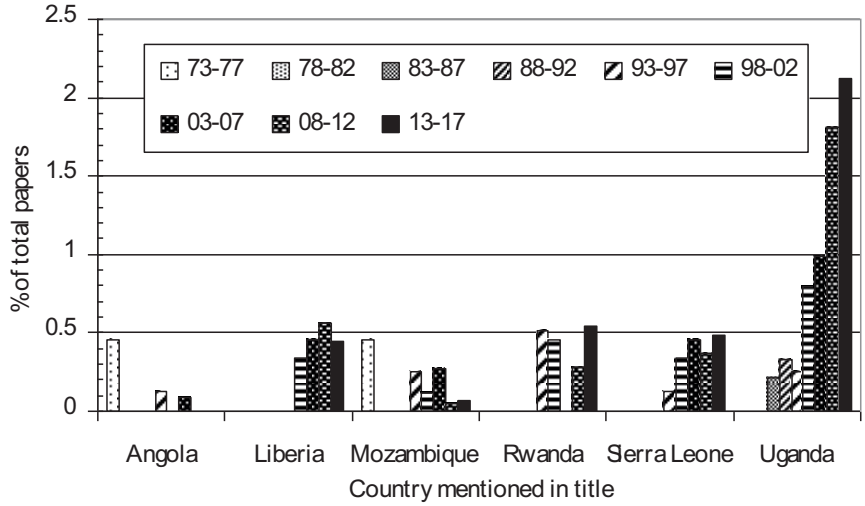


Figure 3. Mentions of African countries in the titles of WARHE papers from 1973-2017.

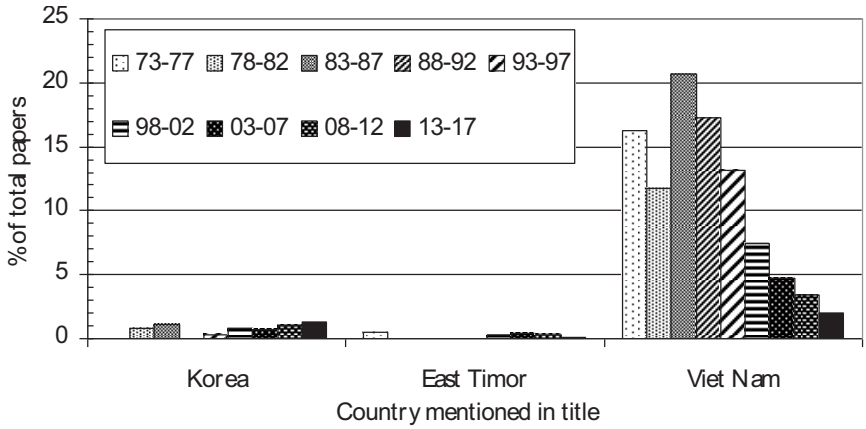


Figure 4. Mentions of East Asian countries in the titles of WARHE papers from 1973-2017.

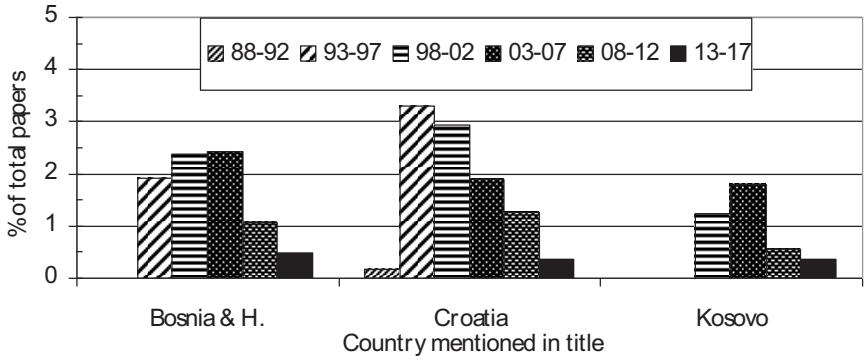


Figure 5. Mentions of Balkan countries in the titles of WARHE papers from 1973-2017.

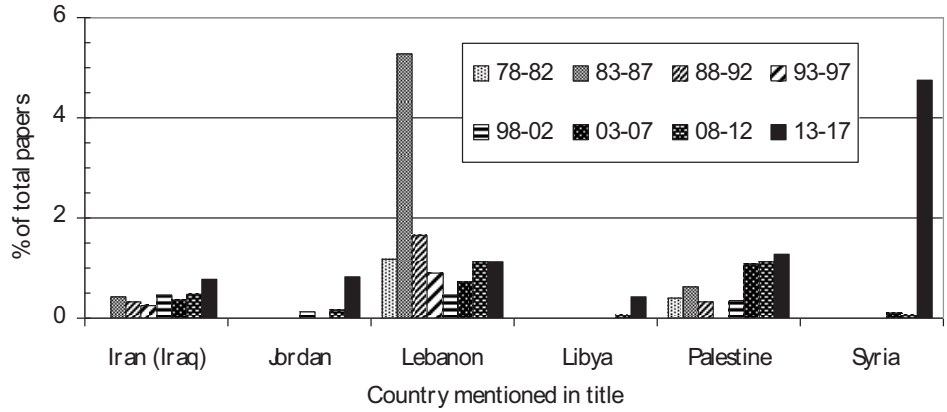


Figure 6. Mentions of some Middle East and North African countries in the titles of WARHE papers from 1973-2017.

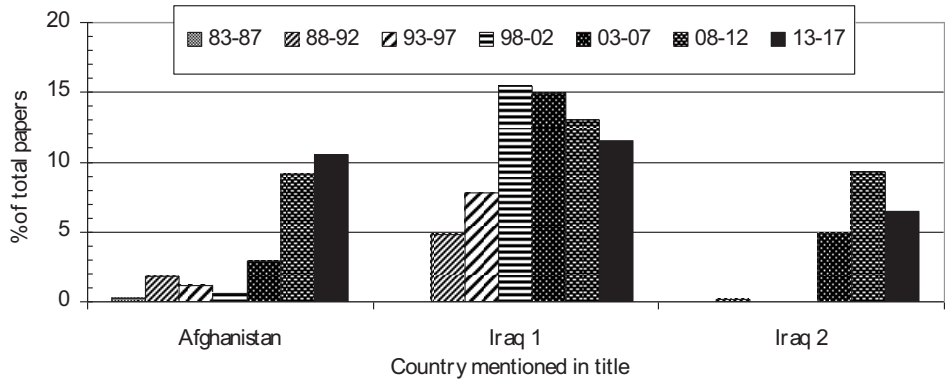


Figure 7. Mentions of Afghanistan and Iraq (two campaigns) in the titles of WARHE papers from 1973-2017.

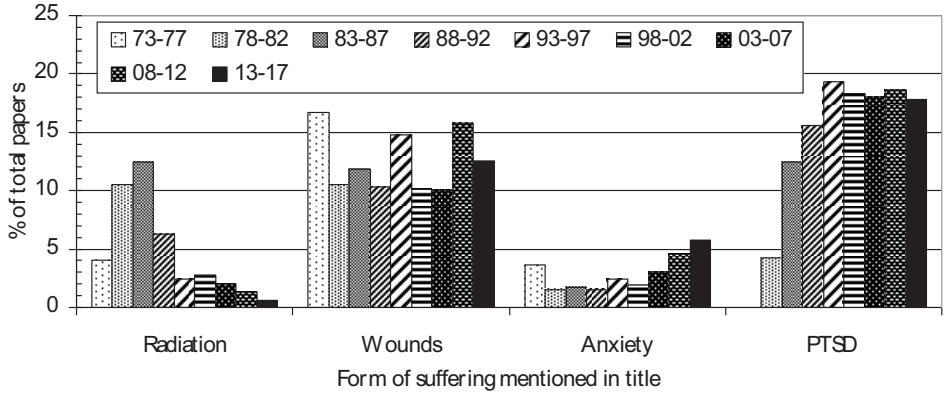


Figure 8. Mentions of types of suffering in the titles of WARHE papers from 1973-2017.

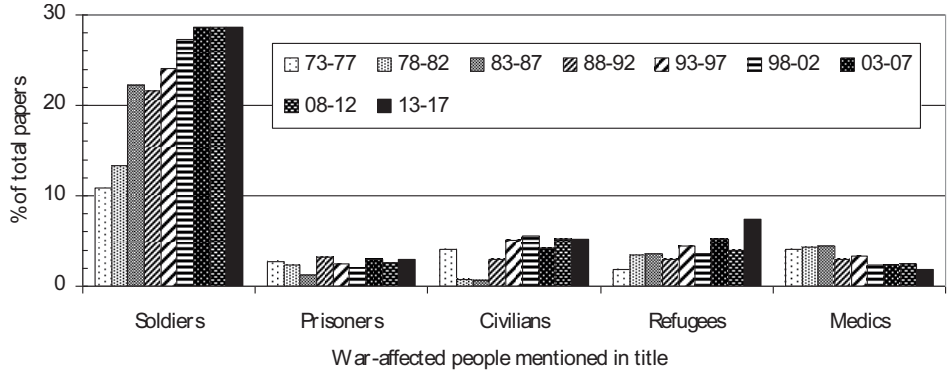


Figure 9. Mentions of the people affected by war in the titles of WARHE papers from 1973-2017.

A cross-tabulation of types of suffering and the people affected showed that for the last 45 years, PTSD was researched on 1631 papers of which 974 concerned military personnel (38%), but fewer of the papers on prisoners (24%), civilians (14%) and refugees (11%). The other main type of suffering, physical injuries, with 1228 papers, was researched almost equally (11%) on soldiers, prisoners and civilians, but not on refugees (only 3%). The effects of physical wounds on medical personnel were, however, more researched than the effects of PTSD.

The journals most frequently used for publication of WARHE papers

The papers in our database were published in 2545 different journals. *Military Medicine* was by far the most used, with 1236 papers (nearly 12% of the total). Most of the other papers were published in general medical journals, but there were four other military medical journals: *Médecine et Armées* (68 papers), *Journal of the Royal Army Medical Corps* (66), *Conflict and Health* (58), and *Military Psychology* (54).

The citations to papers from different countries

The results of our analysis for the leading 11 countries are shown in Table 3, where the countries are ranked by their "Worldscale" value, representing their presence among the top 5% of cited papers with 31 or more cites. However, this ranking is little different from a ranking based on mean ACI values.

Table 3. Citation performance in WARHE papers by the leading 13 countries in terms of fractional count outputs, 1946-2013.

Country	*Citable*	*Citations*	*ACI mean*	*ACI >= 31*	*% in top 5%*	*WS value*
Japan	109.1	1748	16.0	15.7	14.37	278
Netherlands	87.9	905	10.3	7.3	8.33	161
USA	3764	42743	11.4	301	8.00	155
Germany	199	1647	8.3	12.5	6.28	121
UK	500	4508	9.0	27.9	5.58	108
Switzerland	53.3	358	6.7	1.5	2.78	54
Australia	196	1454	7.4	4.7	2.42	47
Canada	122	806	6.6	2.1	1.69	33
Israel	307	1680	5.5	4.8	1.56	30
France	141	390	2.8	2.2	1.54	30
Iran	48.9	184	3.8	0.2	0.34	7

Because of the expansion of the WoS over the years, and particularly in the 21st century, the mean value of ACI has been rising with time from less than 2 in 1973-77 to 6 in 1993-97 and 13 in 2008-13. The overall mean value of 8.74 citations is influenced by the big increase in numbers of WARHE publications in the last 15 years, see Figure 1.

Discussion

Characterisation of the field of literature on warfare and health is a challenging task, given the numerous contributions from a diverse and interdisciplinary network. A limitation of this study is that only one database was used, namely the WoS, and others such as PubMed or Scopus would have generated additional papers (Larsen & von Ins, 2010; Mongeon & Paul-Hus, 2016). Although the final version of the filter was a compromise between the need to maximise both precision and recall, the shortfall of papers on Syria indicates that some additions of individual papers would have been beneficial. [However, since US forces were only peripherally involved, as compared with the wars in Afghanistan and Iraq, it would be expected that there would be fewer papers about this country in the WoS because of its bias in coverage, especially in the social sciences.]

The subject of "warfare and health" is clearly one whose research activities, although relatively small, are *expanding rapidly*, particularly since 2007 (Figure 1). The literature on the subject is dominated by contributions from the USA to an unusual degree, probably because of the large number of wars in which the country has been involved, and also because China, which is currently a major contributor to biomedical research, has not been involved in any recent conflicts (Table 1). The most prominent countries, relative to their biomedical research outputs, are all ones that have seen major conflicts during the last 50 years.

International collaboration is less frequent than for biomedical research, probably because a large number of papers are concerned with the mental and physical health of a country's armed forces (Figure 9), and relatively less with the effects of warfare on civilian populations and the refugees that result from them. Indeed, Figure 9 shows that there are three fifths as many papers in 1973-2017 on the effects on prisoners of war (256) as on the civilian populations (441), and slightly more than there are on medical personnel who take part in deployments (243).

The *country subject areas* of the papers have been changing over the last 45 years as well. They naturally reflect the time-lines of the various conflicts, although they are dominated by US involvement in Viet Nam and the Middle East countries (Figures 3 and 6). The volume of literature on a given conflict does not seem to parallel the duration or human toll. For example, there is a dearth of papers about the profound numbers of civilian casualties (and much greater death toll) in Africa from both civil and anti-colonial wars (Figure 2). There is a paucity of papers examining contemporary conflicts. Despite being the location of one of the greatest humanitarian disasters of the modern era, Yemen has received little research attention, with only five papers. The numbers of papers about Syria (150) also seem rather low (Figure 5). There were an additional 13 papers where Syria was mentioned in the abstract but not in the title; there were also 13 papers with Syria in the title but not in the abstract (for six, no abstract was available in the WoS). So the total might have been about 164 if abstracts for all papers had been available (9893, or 95%, had them in the WoS). However, this study is restricted to research papers and reviews, and does not capture other publication types such as commentaries, letters, news items and editorials. Many of the publications on Syria have been commentaries and calls to action – these are not included in our publication count.

The *health aspects* of war have increasingly turned to the mental, as opposed to the physical ones (Figure 7), although physical wounds accounted for an average of 12% of the papers in

1973-2017. The effects of ionising radiation from warfare are now studied much less than formerly as the threat of nuclear war has receded from the 1960s and 1970s. Anxiety, depression and suicide papers have almost doubled from 2.3% of WARHE papers in 1973-87 to 4.5% in 2003-17, and PTSD papers (some of which also involve the other mental disorders) increased even more from 5.6% to 18% in these time periods. It is clear that the major emphasis of research is on the effects on military personnel rather than on the civilians who suffer from the war (Figure 8), although research on the latter has been increasing slightly as a percentage of the research output.

The *citation analysis* provides an unusual result, namely that the most highly cited papers are those from Japan. This is because of its papers on the effects of ionising radiation, which have been important for the safe development of civil nuclear power and other activities involving radioactivity. The Japanese contribution to this subject area was 97.4 papers out of 311, or 31%, and most of them were written without international collaboration (83 out of 109, or 76%).

Acknowledgement
We are grateful to Philip Roe of Evaluametrics Ltd for the provision of macros that enabled us to carry out the analysis. GL was funded by UK Research and Innovation GCRF Research For Health In Conflict (R4HC-MENA); developing capability, partnerships and research in the Middle and Near East (MENA) ES/P010962/1

References

Chatfield-Ball, C., Boyle, P., Autier, P., *et al*. (2015) Lessons learned from the casualties of war: battlefield medicine and its implication for global trauma care. *Journal of the Royal Society of Medicine*, 108(3), 93-100

Cirillo, V.J. (2008) Two faces of death - fatalities from disease and combat in America's principal wars, 1775 to present. *Perspectives in Biology and Medicine*, 51(1), 121-133.

Dozier, A.M., Martina, C.A., O'Dell, N.L. *et al*. (2014) Identifying Emerging Research Collaborations and Networks: Method Development. *Evaluation & the Health Professions*, 37(1), 19-32

Frost, A., Boyle, P., Autier, P. *et al*. (2017) The effect of explosive remnants of war on global public health: a systematic mixed-studies review using narrative synthesis. *Lancet Public Health*, 2(6), E286-E296

Kim, Y. Lim, H.J. & Lee, S.J. (2014) Applying research collaboration as a new way of measuring research performance in Korean universities. *Scientometrics*, 99(1), 97-115

Larsen, P.O. & von Ins, M. (2010) The rate of growth in scientific publication and the decline in coverage provided by Science Citation Index. *Scientometrics*, 64(3), 575-603

Lewison, G. (2011) Definition of Cancer Research: Journals, Titles, Abstracts or Keywords? *DESIDOC Journal of Library & Information Technology,* 31(5), 333-339

Lewison, G. (1996) The definition of biomedical research subfields with title keywords and application to the analysis of research outputs. *Research Evaluation*, 6(1), 25-36

Lewison, G. & Paraje, G. (2004) The classification of biomedical journals by research level. *Scientometrics*, 60(2), 145-157

Lewison, G., Thornicroft, G., Szmukler, G. & Tansella, M. (2007) Fair assessment of the merits of psychiatric research. *British Journal of Psychiatry*, 190, 314-318.

Liao, C.H. & Yen, H.R. (2010) Quantifying the degree of research collaboration: A comparative study of collaborative measures. *Journal of Informetrics*, 6(1), 27-33

Mongeon, P. & Paul-Hus, A. (2016) The journal coverage of Web of Science and Scopus: a comparative analysis. *Scientometrics*, 106(1), 213-228

Savanur, K. & Srikanth, R. (2006) Modified collaborative coefficient: a new measure for quantifying the degree of research collaboration. *Scientometrics*, 84(2), 365-371

Exploring the borders of a transregional knowledge network. The case of a French research federation in green chemistry[1]

Marion Maisonobe[1] and Bastien Bernela[2]

[1]*marion.maisonobe@ parisgeo.cnrs.fr*

UMR Géographie-cités, CNRS, Université Paris 7, Université Paris 1, Paris (France)

[2] *bastien.bernela@univ-poitiers.fr*

University of Poitiers, CRIEF EA2249, Poitiers, France

Abstract

Competition in research has led to the emergence of numerous consortia of laboratories, designed to improve their participants' visibility. This article aims to understand the determinants of these new structures through the case study of a French federation of laboratories in green chemistry. Working from bibliographic and qualitative data, we examine this federation's geographical and institutional scope, highlighting the importance of i) prior collaborations in French chemistry, ii) interpersonal relations between consortium members, and iii) policy and scientific incentives. This research highlights the role of the territorial scope of the consortium not only as an argument for the consortium to emerge but also as a resource for its development and visibility among academics and industrial partners.

Introduction

In the current context of academic competition, research organizations are attempting to improve their visibility (Yudkevitch *et al.,* 2016). This quest for international profile can be pursued through at least three strategies: being big (size effects associated with the notion that "big is beautiful"), being in the right place (according to the idea that location matters) or being in the right network (strategic alliances and knowledge flows). The former two strategies often work in combination, in a context where scientific concentration policies tend to favour large laboratories based in metropolises (Grossetti *et al.*, 2016). The latter strategy is best adapted for medium-sized networks in non-central locations. Such strategies are often driven by public policies as they serve the purposes of deconcentration and territorial planning in higher education and research.

In this article, we examine the geographic coherence of a cross-regional scientific network in green chemistry supported by public institutions at both the national and the regional level. The network we look at is a network of research laboratories. We hypothesize that such a network is reliant on research practices and in particular on prior or potential scientific collaborations between network members. By adopting a spatial approach, our intention is to better assess the role of this coordination mechanisms and for instance, its ability to fix "network failures" (Lucena and Vicente, 2019). Such "failures" can be the product of long-term spatial processes, or path dependencies, that shape and restrain the organization and evolution of inter-urban scientific networks.

The name of the network we look at is INCREASE, which is a French green chemistry federation. Led by the CNRS[2], it comprises of six laboratories and a club of industrial businesses based in Western France. The first part of the article presents some background information on the French green chemistry sector, as well as the theoretical framework used to investigate the emergence and geographic scope of this research federation. The next section presents the data used for this case study, including both bibliometric data as well as information collected through qualitative interviews with consortium

[1] A longer version of this article in available in preprint: https://hal.archives-ouvertes.fr/hal-02053595
[2] The French National Center for Scientific Research (CNRS)

researchers. In the last section, we present research results. This research allows us to highlight three types of determinants that shape the consortium's geography: structure effects (impact of past activities on the geography of collaborations); network effects (role of pre-existing inter-individual relations); and policy effects (link between policy incentives and cognitive factors).

Studying the structuring of a scientific network in green chemistry

Green chemistry research in France

Along with nano-chemistry, green chemistry is one of the groundbreaking sectors that have over the past fifteen years transformed the way chemistry is taught, organized and practiced (Morris, 2011; Milard and Grossetti, 2017). Although fossil fuels remain the main raw material used by the chemical industry, alternatives to petro-chemistry are the object of a growing number of studies, drawing mounting interest in the industry (Moiseev, 2016). Unlike nano-technology research, which developed in a limited number of scientific hubs due to the spatial concentration of dedicated funding (Bozeman *et al*., 2007; Robinson *et al.*, 2007; Meyer, 2011), green chemistry in France has benefitted from support mechanisms that encouraged the creation of territorialized networks. In this area, no less than four programs have been supporting the clustering of researchers and industries on a geographic basis:

- *Instituts pour la Transition Energétique* (Institutes for the Energy Transition, or ITE), including PIVERT (an institute of excellence on low-carbon energy created in May 2012) and the IFMAS (*Institut Français des Matériaux Agrosourcés*) launched in 2013, both in the North of France;
- *Institut Carnot* 3BCAR, created in 2006 with institutions mainly based in Southern France;
- Competitiveness clusters, including AXELERA, created in 2005 in the *Rhône-Alpes* area which now counts 345 partners (industries and laboratories); the IAR cluster (Industry and Agro-Resources), created in 2007 in the same geographic area as PIVERT (Northern France); and Trimatech, a cluster involving Southern cities including Montpellier, Aix-Marseille and Nimes;
- and finally, the INCREASE Research Federation led by the CNRS, involving several laboratories based in Western France (Table 1), which forms the object of this study.

INCREASE (International Consortium on Eco-conception and Renewable Resources) is a public-private collaborative network that was inaugurated on May 13, 2016. Sponsored by the CNRS, it is dedicated to eco-design and renewable resources, with the aim of developing green chemistry by using biomass – a renewable carbon source – as a raw material. It involves firms from different sectors in the chemical industry (cosmetics, agri-food and detergents) as well as eight academic research centers based in Western France (Poitiers, Rennes, Toulouse, Bordeaux, Nantes and La Rochelle). Another of INCREASE's objectives is to promote the training of young researchers and the dissemination of knowledge in the field of green chemistry through the organization of the International Symposium on Green Chemistry (ISGC), which takes place every two years in La Rochelle.

The INCREASE network is the latest addition to a range of institutional programs. According to its participants, the consortium aims to fill a gap in the geography delineated by prior programs. These programs had shaped two major active areas in green chemistry: the *Rhône-Alpes* area, organized around the city of Lyon, and the North-East area between Amiens and Reims. In addition, laboratories in Toulouse, Bordeaux and Montpellier (Greater South-West region) participated in several large networks such as 3BCAR led by the *Institut Carnot*. This map produced the impression that laboratories in the South, East and North of the country were integrated to large networks, while laboratories based in the West were more isolated. INCREASE associates cities located across four different regions: *Occitanie*, *Nouvelle-Aquitaine*, *Pays de la Loire* and *Brittany*. Although this territory has no administrative

existence as such, it more or less coincides with the area known as the "Atlantic arc". Its territorial cohesion has been the object of historical debates about France territorial development (Brunet, 1993).

Table 1. Presentation of INCREASE members and their respective urban areas.

Laboratories				*Urban areas*			
Names	*Number of permanent staff*	*Number of INCREASE members*	*Scientific contribution to INCREASE*	*Lab location*	*French rank in urban population*	*French rank in scientific production*[3]	*French rank in chemistry*[4]
IC2MP	150	7	Catalytic conversion of biomass to fine chemicals.	Poitiers	41st	20th	16th
ISCR	300	32	Organometallic homogeneous catalysis and molecular valorization.	Rennes	10th	9th	8th
LGC	160	24	Chemical engineering and process chemistry.	Toulouse	4th	4th	4th
LCPO	40	10	Synthesis of polymers and development of functional polymeric materials.	Bordeaux	5th	7th	6th
ISM	130	2	Life-cycle assessment.				
BIA-INRA	190	2	Biopolymers within the biomass and physical chemistry.	Nantes	8th	13th	12th
LIENSs	100	21	Assessment of the chemical contamination of coastal areas.	La Rochelle	48th	39th	39th

This coastline running from Brittany to the Greater South-West region is dotted with medium-sized cities of limited influence including Poitiers, which hosts the headquarters of the INCREASE Federation and has the region's oldest university. Poitiers is located near several other medium-sized cities with more recently established universities: in the sixties for Angers, Limoges and Orléans, and in the nineties for La Rochelle (Milard and Grossetti, 2017). The resulting network of higher education and research institutions is relatively dense, with a positive scientific momentum, in particular in chemistry, both in terms of output, diversity and visibility (Milard, 2012).

While the leadership of chemistry researchers in this area is unquestionably an asset, this does not suffice to conclude that the territory delineated by the boundaries of the INCREASE consortium is necessarily characterized by an outstanding scientific cohesion in chemistry, or for that matter in green chemistry. Actually, the geographic scope of the INCREASE consortium appears to be, at least in part, the product of preexisting relations between its members. As explained by Harrisson (2016), it is interesting to understand how their founders set the boundaries of such "constellations", and to identify the criteria that determine whether a site is integrated into the consortium or not.

Framework and hypotheses

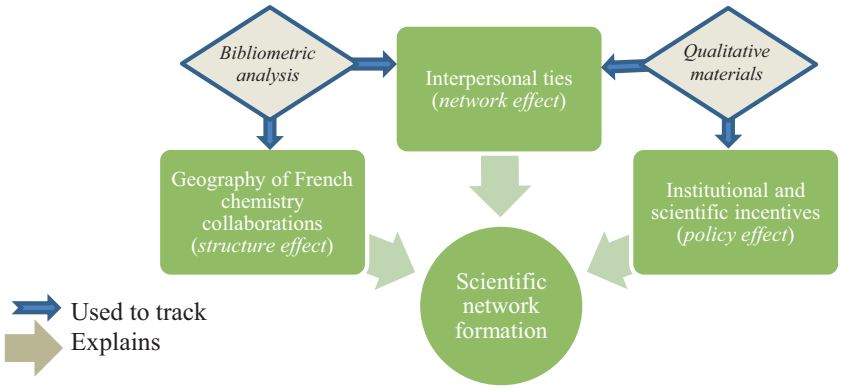
Since the 1990s, literature on science and innovation has focused either on synergy effects between science and innovation stakeholders on a city-wide or regional scale (Cooke *et al*., 1997; Asheim and Coenen, 2005; Catini *et al*., 2015), or on complementarity effects achieved through the networking of clusters, cities and regions on a continental or global scale (Bathelt *et al*. 2004; Sebestian and Varga, 2012; Morrisson *et al*., 2013). In this literature, exchanges of knowledge that are not intra-local or intra-regional are categorized as "non-local links" (Meyer *et al*., 2011) and their geographic coherence is hardly ever discussed, except by analyzing complementarities between "distance effects" and "network effects" (Bergman and Maier, 2009; Ter Wal and Boschma, 2009; d'Amore *et al*., 2013). A few studies suggest however that a more in-depth study of the geography of scientific exchanges can reveal interesting patterns: for instance, the decline of border effects in the organization of European exchanges

[3] Number of publications across all disciplines in the SCI Expanded in 2013
[4] Number of publications in chemistry in the SCI Expanded in 2013

(Hoeckman *et al.*, 2010); the enduring role of the national scale in the global organization of scientific activity (Maisonobe *et al.*, 2016); or the importance of geographic proximity for science-industry partnerships (Ponds, 2007). Our research on the inter-urban relations that are driven or strengthened by a French green chemistry network of laboratories belongs to this latter line of research.

We developed an analytical framework to examine the scientific cohesion of the territory formed by the INCREASE federation and to understand the determinants that shaped its emergence, looking at different levels of interpretation. In line with the analytical framework presented by Gulati and Gargiulo (1999), we worked from the idea that the embeddedness of relations is decisive in understanding the emergence of inter-organizational networks. To investigate the specific case of a cross-regional research and innovation network, our framework considers three levels of explanation (Figure 1).

Figure 1. Framework of scientific network formation



First, to understand the context in which INCREASE emerged, we examine the French map of chemistry – and in particular, special relations in this field. Drawing on the concept of "embeddedness" (Granovetter, 1985; Grossetti, 2008), we consider that the network of scientific collaborations in chemistry that existed prior to the creation of the Federation partly explains the decision of federating laboratories in Western France, as well as the fact that this decision originated from Poitiers researchers. Using a map showing the national organization of scientific activity, we can test the hypothesis that there was indeed a need for a better coordination of laboratories in this part of France. For this purpose, we carried out a bibliometric analysis of scientific collaboration networks between French cities. Along with Katz (1994), we consider that relations of scientific collaboration can be assessed from data on the co-authoring of scientific articles.

H1: The state of scientific collaborations in chemistry between French cities shows a lack of coordination in the "Atlantic arc", which justifies the federation's creation (structure effect).

Secondly, we focus on interpersonal relations that existed prior to the creation of the consortium, and that could explain its emergence. The presumed need for cross-regional coordination can only be fulfilled if there is at least some level of acquaintance between the parties involved in the emerging network. To assess these relations, we used data collected through interviews with federation members and through data on these members' scientific collaboration networks, which we extrapolated from a corpus of bibliographic data from the *Web of Science* (the contents of this corpus are detailed later).

H2: The federation's geographic scope derives from prior relations of acquaintance and from projects of future collaborations between its members (network effect)

Finally, although the federation is a "bottom-up" construct, its existence depended on it being signed off by public authorities. As a matter of fact, the consortium would not have any official existence without the CNRS's moral and financial backing. This means that its scientific agenda is in line with the CNRS's scientific and organizational objectives. The consortium benefitted from a context that made it possible, acceptable and even desirable in the eyes of CNRS leadership. These circumstances are not just the product of the CNRS's internal policy, but also of a range of political, scientific and socio-environmental priorities on various scales (local, regional, national, European). To examine the policy level, we used qualitative materials including interviews, reports and grey literature.

H3: The consortium's emergence was made possible by its alignment with a number of scientific and political priorities (policy effect).

By considering these various levels of causality, as represented in Figure 1, we can identify the social, scientific and historical factors that led to the integration of specific members and organizations into the consortium, the choice of its geographic scope and therefore its geography.

Data collection

Capturing prior networks through a bibliometric analysis

The bibliometric corpuses analyzed in this article are from the *SCI Expanded*. We consider three corpuses of publications:

- The first corpus (*corpus A:* 313,165 publications) includes all French contributions to journals listed in the *SCI Expanded*, regardless of their discipline (see Table 1):
- The second corpus (*corpus B:* 37,790 publications) includes all French contributions to publications listed as chemistry journals in the *SCI Expanded*:
- The third corpus (*corpus C:* 1,813 publications*)* includes all contributions published by consortium members. To delineate this corpus, we asked the representatives of each laboratory to give us a list of all the members whose current work is consistent with the research agenda of INCREASE. As a result, we obtained a list of 98 INCREASE researchers, unevenly distributed across the 7 labs (see Table 1). This unevenness can be explained both by the diverse sizes of the labs and by the diverse views team leaders have of the INCREASE research agenda.

For all three corpuses, we considered publications from 2007 to 2014. 2014 was chosen as a point of reference because we started this research in 2017 and it usually takes between two and three years for information on more recent publications to be up-to-date. 2007 was chosen as a starting point because a large number of initiatives aimed at structuring green chemistry in France were delivered after this date (Schultz, 2016). Those include in particular the launch of a new call for projects, and the creation by the CNRS of a national network of green chemistry researchers (CPDD for Chemistry and Processes for Sustainable Development).

To obtain a global overview of these datasets, we used the spatial bibliometric method developed by one of us in collaboration with a team of geographers and sociologists to process the entire contents of the *SCI Expanded* on the scale of urban areas (Maisonobe *et al.*, 2016). Urban areas were chosen as our resolution level because the administrative boundaries of cities are not comparable, and scopes of universities and labs are very heterogeneous from a place to the next. Knowing that the majority of co-authoring links happen within a single urban area, we consider interesting to focus on co-authoring links that develop between different urban areas. These links reflect collaborations that contribute to structuring scientific communities beyond the local scale (regional, cross-regional, national, macro-regional and global scale). We fraction the links' weight according to the number of distinct urban areas that contribute to each co-publication ("*Whole Normalized Counting"*).

Understanding the emergence of the Federation through stakeholder interviews

In order to understand the set-up of the INCREASE federation and to qualify the relations between INCREASE members as well as their integration into national and international green chemistry networks, we conducted two stages of interviews.

The first stage focused on INCREASE stakeholders: seven research lab representatives, three industry leaders as well as three institutional partners (University of Poitiers, Poitiers' Regional Council and the CNRS's chemistry department). 13 interviews were completed in 2016, of an average duration of 90 minutes. Their objectives were to question the stakeholders on the consortium's genesis, history and objectives, as well as on the benefits derived by each organization from participating in the network.

The second stage took place during the ISCG (International Symposium on Green Chemistry) organized by INCREASE in La Rochelle in May 2017. At the end of this week-long conference which brought together over 800 participants from 48 different countries, we completed 20 interviews of an average duration of 20 minutes, including 7 with academic members of INCREASE (who had not been interviewed during stage 1), and 8 with researchers or industrials who had co-authored a publication with an INCREASE member. We conducted these interviews with the help of data from the bibliometric analysis of *Corpus C*. Interviewees were asked questions on their co-authoring networks and on their participation in national and international green chemistry networks.

Results

The organization of the chemistry field in France

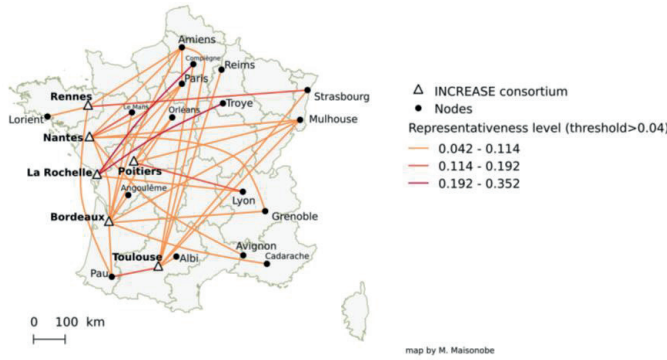
For many stakeholders in the French chemistry sector, the choice of federating laboratories based in the West of the country was coherent with the discipline's national structure. While laboratories in the Eastern half of France appeared well connected and inserted into diverse networks, those based in the West had little interaction between them in spite of their visibility and notoriety in the field of green chemistry. According to the founder of INCREASE located in Poitiers, whom was interviewed shortly after the launch of the federation in 2016, the French map of green chemistry networks was organized as followed: to the North-East, "*a giant banana-shaped area*" spanning from Lille to Reims (the PIVERT institute); to the South-East, "*Lyon's chemical industry hub, with the Axelera competitiveness cluster*"; and further South, Montpellier with the 3BCAR *Carnot Institute*, and Toulouse with the TWB (Toulouse White Biotechnologies); "*but the whole of the Greater West was left out, with isolated laboratories*". For this reason, the idea of structuring this field in the "Atlantic arc" emerged, which led to the creation of INCREASE. To confirm this interpretation and understand the national context in which this research federation emerged, we studied empirical data on the co-authoring of chemistry papers. Our analysis embraced urban areas, measuring the number of co-publications associating two different urban areas over the 2007-2014 period, just before the launch of the INCREASE consortium.

In order to qualify the existing relations between the six cities involved in the INCREASE chemistry consortium, we compare the spatial distribution of cross-urban collaborations in chemistry and across all disciplines. Suppose that x_{ij} is the number of co-authored papers between urban areas *i* and *j* in chemistry (corpus B), and y_{ij} the number of co-authored papers between urban areas *i* and *j* across all disciplines (corpus A). In order to compare x_{ij} with y_{ij}, we normalize them according to the following formula[5]: $X_{ij}=\frac{x_{ij}}{x_{i\cdot}+x_{\cdot j}}$ for corpus B, and $Y_{ij}=\frac{y_{ij}}{y_{i\cdot}+y_{\cdot j}}$ for corpus A. Then, we measure the deviation between X_{ij} and Y_{ij} using a measure of deviation from independence, which indicates the degree of

[5] As noted by van Eck and Waltman (2009), there are various ways of normalizing collaboration data. In this article, we chose to normalize values based on the sum of weighted degrees for each city involved in a collaboration.

representativeness of co-authoring links between two cities in chemistry compared to in all disciplines. This indicator of representativeness' formula is: $p_{ij} = \frac{X_{ij} - Y_{ij}}{\sqrt{Y_{ij}}}$. If p_{ij} is greater than zero, then the intensity of relations observed in the field of chemistry is greater than expected in light of scientific collaborations between the two cities across all disciplines. Figure 3 shows overrepresented relations. This map indicates that the chemical industry hub around Lyon accounts for a relatively small share of relations with INCREASE cities. Only researchers from Poitiers and La Rochelle have a special relationship with Lyon in chemistry. Aside from a few exceptions, the main collaborations of the six INCREASE cities in chemistry are mostly located in the North and West of France. In particular, INCREASE cities share strong links with small and medium-size western cities: Pau, Orléans, Le Mans, Angoulême, Albi, and Lorient. Figure 3 thus suggests an East-West divide confirmed by the interviews.

Figure 3. Privileged co-publication links with INCREASE urban areas in chemistry versus all disciplines combined



Privileged co-publication linkages of INCREASE urban areas in chemistry versus all disciplines combined (2007-2014). Only the cities with at least 5 collaborations with at least one of the six cities belonging to the INCREASE consortium are included.
Source: Web of Science (articles, reviews, letters)

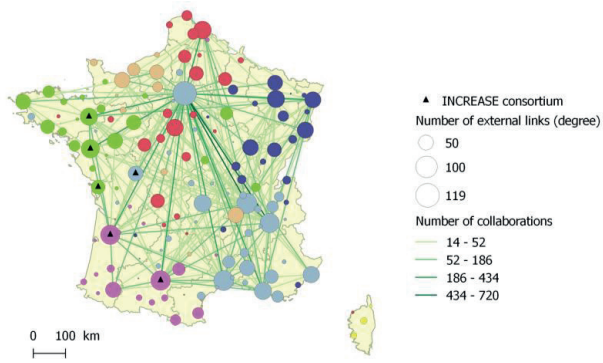
According to our interviewees, aside from their relative independence from the "chemistry valley" (South-East France), Western cities are traditionally isolated from each other. In order to ascertain this second fact, we apply a community detection algorithm to the network of collaborations between French cities in chemistry. This algorithm partitions the network into several groups characterized by a greater density of interactions between their members than with external groups. By applying the "Louvain" method[6], we identify seven collaboration areas in chemistry represented in Figure 4.

We observe that INCREASE cities belong to different groups. Rennes, Nantes and La Rochelle connect with the Brittany group as well as with a few medium-sized cities in the center of France (green group). Bordeaux and Toulouse are part of a group that includes the Greater South-West (purple group). Located at the intersection of Brittany and Greater South-West, Poitiers is isolated and connected to the central network formed by Paris and Lyon (light blue group). The map shows that Poitiers sits in a unique position. The choice of creating a Poitiers-based consortium thus appears as an ambitious one, as this initiative could transform established networks. Rather than being a satellite of Paris and Lyon, Poitiers aspires to become a bridge in the Atlantic arc. According to the consortium's founder, Poitiers indeed holds a strategic position in this network of six cities: "*When you look at the map, Poitiers is at the center, so when we have meetings everyone is at about two hours' distance, which is pretty good for our operations. Proximity is important – you can always use videoconferences but they're not the same.*"

[6] This method was applied using default resolution parameters (1). The result thus obtained is rather similar to that obtained with other clustering methods. In all cases, cities in Brittany and the South-West are in distinct groups.

This bottom-up decision, signed off by national CNRS leaders, intends to "fix" the lack of connections between the North and South of the Atlantic arc in chemistry, while deeply transforming Poitiers' position in the overall structure (Lucena and Vicente, 2019).

Figure 4. Community structure analysis of collaboration networks between French urban areas in chemistry.



Community structure analysis of the collaboration network between French urban areas in chemistry (2007-2014). Colored groups are Louvain communities: a community is a group of nodes more connected to each other than to the other groups.
Only the cities with at least 10 publications co-authored between 2007 and 2014 in chemistry are included.
Source: Web of Science (articles, reviews, letters)

map by M. Maisonobe

The role of interpersonal relations

In the absence of much prior intensive scientific collaboration between network members (only Toulouse-Rennes; Nantes-La Rochelle, Nantes-Poitiers and Bordeaux-Toulouse had some prior co-authoring linkages according to Corpus C), other types of relations played a crucial role in the design and roll-out of the INCREASE network. In our interviews, we noted that several members of the consortium knew each other, or were at least aware of each other, before the federation's launch. These acquaintance networks were decisive in the partners agreeing to join INCREASE when contacted by Poitiers' laboratory. The interviews allow us to distinguish between three types of relations: ties formed through researchers' professional mobility experiences; ties resulting from teaching and training functions (lectureships, thesis juries, etc.); and finally, ties formed between researchers contributing to common institutions, conferences and research programs.

This confirms the importance of networking in the academia organized as a professional system. Relations can form over the course of an individual's career, including between laboratories that do not have a record of scientific collaboration. For instance, the Rennes and Poitiers INCREASE representatives met when they were both doing a PhD in Dijon, on subjects with no relation to green chemistry. Furthermore, the federation's founder briefly worked in Rennes before taking up his post in Poitiers. For this reason, although the Rennes and Poitiers laboratories have never as yet collaborated on green chemistry projects, they will be able to do so in the future thanks to their relation of trust and to the consortium's incentive mechanisms.

Similarly, although the relation between Poitiers and La Rochelle had not produced any scientific collaborations prior to the federation's launch, this relation exists – if only because both cities used to belong, unlike the other four, to the same administrative region. The INCREASE representative in La Rochelle explains: "*I've never published with anyone from Poitiers* (…) *I've never published with anyone from IC2MP* [Poitiers research lab]. *But I do keep in touch with them: I take part in their recruitment days, I teach in a training course we run in common and we were in contact at the time of*

the creation of the Green Chemistry Institute." The Poitiers laboratory has indeed long specialized in green chemistry, and has done so even before the arrival of the INCREASE co-ordinator in Poitiers. An initiative led by the previous generation of researchers had helped position the Poitiers lab in the national CPDD program (2007-2009), and federate local researchers with support from the Region.

Aside from those who get to know each other through working or meeting in the same laboratory at a point in their career, or through contributing to common training programs or thesis juries, others are well known to others because of their prominence in the field. For instance, two INCREASE researchers explained knowing each other before because they had been members of the same regional branch of the *Société Chimique de France*. This is also the case with the Bordeaux and Toulouse INCREASE representatives, known for having sat on the CNRS's National Committee and on the National Research Agency (ANR), an agency in charge of the evaluation and selection of calls for projects. After 2007, some programs led by the CNRS and the ANR contributed to the national structuring of the "sustainable" chemistry field. These programs played an essential part on three levels: they helped the teams working on this subject to become aware of each other; they helped foster links with industrials; and they contributed to setting a scientific agenda.

Policies supporting the emergence of a regional multi-skill network

The networking of research institutions, as observed in this study through the creation of INCREASE, can be motivated by purely scientific motives: the network is often seen as a way of solving research problems by cross-pollinating skills and knowledge from fields with little communication between them. The development of green chemistry is a good illustration of this need for cross-pollination: today's challenges require connections between chemistry (reaction, catalysis), processes (chemical engineering), biomass issues (biology) and lifecycle analysis (environmental impact).

From 2006, the CNRS and the ANR contributed to promoting sustainable chemistry's development and organization in France (Schultz, 2016). In 2006, the CNRS launched a cross-disciplinary program (CPDD), aimed at drawing an inventory of sustainable chemistry teams in France. The CPDD program was led by three laboratories that later formed the core of the INCREASE consortium: Poitiers, Nantes and Toulouse. This initiative offered little funding but created a momentum amongst CNRS researchers. The Toulouse INCREASE representative reported that this program had given him the opportunity to launch a thematic summer school and an international conference on sustainable process engineering. Shortly after this, the ANR launched a large-scale call for projects: "Chemistry and Processes for Sustainable Development", targeting the entire French community, beyond CNRS researchers. This time, some funding was available to support selected projects and the program offered strong incentives for multi-partner projects. The first phase was successful but its outcomes in terms of building links with the industry were disappointing. A second phase was launched in 2010 under the title: "Sustainable Chemistry – Industries, Innovation". The program supported research projects with the ability to produce technology transfers that could benefit the industry, and focused on two aspects: improving communication between chemistry, processes and biotechnology; and assessing the costs and environmental impacts associated with resources, reactions and processes (Schultz, 2016).

These incentives significantly oriented the selection of labs contacted for the creation of the INCREASE consortium. The issue of environmental impacts was first built into the agenda of the ISGC international congress, organized bi-annually since 2013 by the Poitiers lab. Individuals who expressed their interest in this aspect of the congress were then invited by the founder of INCREASE to join the consortium. By involving chemists from Rennes, Nantes and Poitiers, the consortium also brings together specialists on all three types of catalysis, which are hardly ever associated (enzymatic, homogeneous and heterogeneous). This diversity of research subjects is the reason why the various teams had never had the opportunity to collaborate much before. Combining diverse specialties has both scientific and

economic benefits: the joint objectives are to further sustainable chemistry *and* to create a "toolbox" that industrials can access by directly contracting with the consortium rather than having to negotiate with each separate team.

The interviews produce the impression that the INCREASE consortium's geographic coherence is more the result of multiple decisions than that of a clear intention to stay within the boundaries of a pre-defined geographic perimeter. According to the consortium's leader, geographic proximity was not initially perceived as a must. It was even envisaged for the consortium to be "international" from the outset and to include foreign laboratories. Some interviewees explained that the Lille team could also have been part of the consortium due to its prior connections with Poitiers and Rennes, to its scientific specialism in characterization and to its international reputation. Similarly, Montpellier has a significant collaboration record with Bordeaux in the field of polymers and could legitimately have been part of the consortium. However, geography appears to have gradually become a resource on different levels.

Indeed, the consortium's geographic coherence is a strong asset to rally support from national and international institutions and from industry partners. This coherence makes the French map of green chemistry more readable. It also contributes to the international visibility of Western France. In a context of growing international competition with the emergence of new congresses on this theme in Germany and New-Zealand, it appeared relevant to anchor the Poitiers group into a structured local network with an established reputation for its diverse skill base. In terms of science-industry transfers, this geographic coherence is also an asset. Overall, the emergence of INCREASE responds to a trend that is broadly encouraged by today's public authorities and policies: clustering and/or networking research institutions located within one same area that offer a combination of diverse resources and skills. This response to existing needs is driven by the founder of INCREASE, who wants to develop his leadership in a context where the CNRS is keen to train professional research executives. Ultimately, the creation of INCREASE owes as much to his skills as a researcher as it does to those as an academic entrepreneur (Jain *et al*., 2009; Perkmann *et al.,* 2013): an entrepreneur who took advantage of geography to build a consortium that could in turn transform the geography of connections.

Conclusion

Green chemistry refers to a drive in chemistry towards adopting more environmentally friendly practices (Marion *et al*., 2017). These practices, which can be expected in term to spread across all of chemistry, are today becoming structured as a specialism – or at least, as a cross-disciplinary area of work. Its development parallels that of nano-chemistry, which has also contributed to transforming the discipline. In both areas, change does not only affect major international hubs. The emergence of new perspectives in chemistry makes it necessary to move beyond old collaboration habits and to build links between teams that had not been accustomed to collaborate in the past. Coordination mechanisms, such as national or European calls for research projects, were created for this purpose. Another structuring initiative in France has been the opportunity to form consortia of laboratories around specific themes. This opportunity, although rarely taken up, is one of the mechanisms developed by the CNRS to structure the national map of scientific activities. It can facilitate the emergence of cross-regional networks whose boundaries are original in that they do not match existing administrative units – such as for instance the INCREASE consortium, which forms the object of this study.

This article had a three-fold objective: better understanding the creation of the INCREASE cross-regional thematic network; examining the level of scientific structuring of the cross-regional space shaped by this network, or "Atlantic arc"; and studying the determinants of its delineation and of its scientific and geographic scope. For this purpose, we examined the recent map of scientific collaborations in chemistry (structure effect – H1), the various types of existing relations between consortium members (network effect – H2) as well as the impact of policy incentives, scientific and

socioeconomic issues on the scope of this inter-organizational creation (policy effects – H3). These three levels of explanation shed an interesting light on the structuration processes that affected this cluster of Western France cities in green chemistry.

Unlike the alliances studied by Harrisson (2016) initiated by the UK government to facilitate the allocation of higher education and research funding and let alliance members distribute the funds between themselves in proportion with their weight or evaluation scores, the instrument we studied is a flexible one, born from the scientists' desire to federate around a common theme. Member laboratories retain their full autonomy and freedom to work with any other laboratory[7]. Some members may question the consortium's boundaries and wish that such or such laboratory had been included; others are not very convinced by the need for geographic coherence (as discussed above, this coherence is a resource more than a necessity); but all understand the benefits of a consortium of this type.

If a lesson had to be drawn from this study, it would be that it is not for geography to constrain public policies, but instead for public policies to facilitate the activation of connections on different geographic scales. No single level should be privileged above others – for instance through pooling all efforts into the development of local synergies (as was the case with the development of "technopoles"), or into the consolidation of international links. On the contrary, the INCREASE case shows that cross-regional links can also be productive. As shown by recent studies on "network failures" (Lucena and Vicente, 2019), scientific collaboration policies should not set constraints but instead ambition to support researchers and industrials. This involves giving them the instruments they need to facilitate connections that appear to have potential but that are challenging to implement for historical reasons to do with collaboration habits and institutional boundaries (such as the administrative unit of the region). The creation of the INCREASE network was made possible by the flexible model of the "research federation" proposed by the CNRS, which appears particularly adapted to this type of approach.

Acknowledgments

We thank OST-HCERES for providing us the bibliometric data we needed for this research. We also thank Béatrice Milard, Marie Ferru, Michel Grossetti and Olivier Bouba Olga for their comments and advices, as well as the interviews for their precious time and interest for the outcomes of this research. This research enjoyed the funding and material support of the INCREASE federation (FR3707).

References

Asheim, B. T., & Coenen, L. (2005). Knowledge bases and regional innovation systems: Comparing Nordic clusters. *Research Policy*, 34(8), 1173–1190.

Bathelt, H., Malmberg, A., & Maskell, P. (2004). Clusters and knowledge: local buzz, global pipelines and the process of knowledge creation. *Progress in Human Geography*, 28(1), 31–56.

Bergman, E. M., & Maier, G. (2009). Network central: regional positioning for innovative advantage. *The Annals of Regional Science*, 43(3), 615–644.

Bozeman, B., Larédo, P., & Mangematin, V. (2007). Understanding the emergence and deployment of 'nano' S&T. *Research Policy*, 36(6), 807–812.

Brunet, R. (1993). L'Arc Atlantique, son reflet et ses mirages. *Norois*, 40(157), 19–34.

Catini, R., Karamshuk, D., Penner, O., & Riccaboni, M. (2015). Identifying geographic clusters: A network analytic approach. *Research Policy*, 44(9), 1749–1762.

Cooke, P., Gomez Uranga, M., & Etxebarria, G. (1997). Regional innovation systems: Institutional and organisational dimensions. *Research Policy*, 26(4–5), 475–491.

D'Amore, R., Iorio, R., Labory, S., & Stawinoga, A. (2013). Research Collaboration Networks in Biotechnology: Exploring the Trade-Off Between Institutional and Geographic Distances. *Industry and Innovation*, 20(3), 261–276.

[7] If a collaboration is developed between two of the seven labs involved in the consortium, then this collaboration can benefit from the federation's support in the form of a PhD grant, internship, industrial contract, etc.

Granovetter, M. (1985). Economic Action and Social Structure: The Problem of Embeddedness. *American Journal of Sociology*, 91(3), 481–510.

Grossetti, M. (2008). Proximities and Embedding Effects. *European Planning Studies*, 16(5), 629–642.

Grossetti, M., Eckert, D., Maisonobe, M., & Tallec, J. (2016). Four commonly held beliefs about the geography of scientific activities. *Handbook on the Geographies of Innovation*, 223.

Gulati, R., & Gargiulo, M. (1999). Where Do Interorganizational Networks Come From? *American Journal of Sociology*, 104(5), 1439–1493.

Harrison, J., Smith, D., & Kinton, C. (2016). New institutional geographies of higher education: The rise of transregional university alliances. *Environment & Planning A*, 48(5), 910–936.

Hoekman, J., Frenken, K., & Tijssen, R. J. W. (2010). Research collaboration at a distance: Changing spatial patterns of scientific collaboration within Europe. *Research Policy*, 39(5), 662–673.

Jain, S., George, G., & Maltarich, M. (2009). Academics or entrepreneurs? Investigating role identity modification of university scientists involved in commercialization activity. *Research policy*, 38(6), 922-935.

Katz, J. S. (1994). Geographical proximity and scientific collaboration. *Scientometrics*, 31(1), 31–43.

Lucena Piquero, D., & Vicente, J. (2019). The Visible Hand of Cluster Policy Makers: An Analysis of Aerospace Valley (2006-2015) Using a Place-Based Network Methodology. *Research Policy*, 48(3).

Maisonobe, M., Eckert, D., Grossetti, M., Jégou, L., & Milard, B. (2016). The world network of scientific collaborations between cities: domestic or international dynamics? *Informetrics*, 10, 1025–1036.

Marion, P., Bernela, B., Piccirilli, A., Estrine, B., Patouillard, N., Guilbot, J., & Jérôme, F. (2017). Sustainable chemistry: how to produce better and more from less? *Green Chemistry*, 19(21), 4973-4989.

Meyer, M., Libaers, D., & Park, J.-H. (2011). The Emergence of Novel Science-related Fields: Regional or Technological Patterns? Exploration and Exploitation in United Kingdom Nanotechnology. *Regional Studies*, 45(7), 935–959.

Milard, B. (2012). Emergence, internationalisation et autonomisation de villes scientifiques secondaires. L'exemple de la chimie à Nantes, Angers et Le Mans. In *Université et Territoires* (pp. 89–104). Rennes: Presses Universitaires de Rennes.

Milard, B., & Grossetti, M. (2019). Excellence everywhere? Regional development of French scientific output and visibility. *Regional Studies*, 1-11.

Moiseev, I. I. (2016). Biotechnology is storming the heights of petrochemistry. *Kinetics and Catalysis*, 57(4), 405–421.

Morris, P. J. (2011). The Fall and Rise of the History of Recent Chemistry. *Ambix*, 58(3), 238–256.

Morrison, A., Rabellotti, R., & Zirulia, L. (2013). When Do Global Pipelines Enhance the Diffusion of Knowledge in Clusters? *Economic Geography*, (89), 77–96.

Perkmann, M., Tartari, V., McKelvey, M., Autio, E., Broström, A., D'Este, P., ... & Krabel, S. (2013). Academic engagement and commercialisation: A review of the literature on university–industry relations. *Research Policy*, 42(2), 423-442.

Ponds, R., Van Oort, F., & Frenken, K. (2007). The geographical and institutional proximity of research collaboration*. *Papers in Regional Science*, 86(3), 423–443.

Robinson, D. K. R., Rip, A., & Mangematin, V. (2007). Technological agglomeration and the emergence of clusters and networks in nanotechnology. *Research Policy*, 36(6), 871–879.

Schultz, E. (2016). Construire une économie de la recherche sur projets. L'installation de l'Agence Nationale de la Recherche en France et ses conséquences dans les domaines de la génomique végétale et de la chimie durable (Thesis in Sociology). Paris 4, Paris.

Sebestyén, T., & Varga, A. (2013). Research productivity and the quality of interregional knowledge networks. *The Annals of Regional Science*, 51(1), 155–189.

Ter Wal, A. L. J., & Boschma, R. A. (2009). Applying social network analysis in economic geography: framing some key analytic issues. *The Annals of Regional Science*, (43), 739–756.

van Eck, N. J., & Waltman, L. (2009). How to Normalize Co-occurrence Data? An Analysis of Some Well-Known Similarity Measures. *Journal of the American Society for Information Science and Technology*, 60(8), 1635–1651. https://doi.org/10.1002/asi.21075

Yudkevich, M., Altbach, P. G., & Rumbley, L. E. (Eds.). (2016). *The global academic rankings game: Changing institutional policy, practice, and academic life*. Routledge.

The correlation between the level of internationalization of a country's scientific production and that of relevant citing publications

Giovanni Abramo[1], Ciriaco Andrea D'Angelo[2] and Flavia Di Costa[3]

[1] *giovanni.abramo@uniroma2.it*
Laboratory for Studies in Research Evaluation, Institute for System Analysis and Computer Science (IASI-CNR), National Research Council of Italy (Italy)

[2] *dangelo@dii.uniroma2.it*
Department of Engineering and Management, University of Rome "Tor Vergata" (Italy)

[3] *flavia.dicosta@gmail.com*
Research Value s.r.l. (Italy)

Abstract

The growing complexity of scientific challenges demands increasingly intense research collaboration, both domestic and international. The resulting trend affects not only the modes of producing new knowledge, but also the way it is disseminated within scientific communities. This paper analyses the relationship between the "level of internationalization" of a country's scientific production and that of the relevant citing publications. The field of observation consists of Italian publications over the period 2010-2012 and the relative citations accumulated as of 31/05/2017. Findings show: i) the probability of being cited increases with the level of internationalization of the research team; ii) totally domestic research teams tend to cite domestic publications to a greater extent; iii) vice versa, publications resulting from international collaborations involving a minority Italian co-authors tend to be more cited by publications with a totally foreign byline. These results emerge both overall and at the level of individual disciplines.

Introduction

The importance of citation analysis in science is witnessed in the increasingly routine use of citational data to measure the scholarly impact of publications and scientific journals, as well as the research strengths of countries, institutions, departments, and individual scientists.

The foundations of citation analysis lie in normative theory, which posits that scientists cite papers to recognize their influence (Kaplan, 1965; Merton, 1973; Martin & Irvine, 1983). The social constructivist theory objects to this view, claiming that citing to give credit is the exception, while persuasion is the main motivation (Mulkay, 1976; Bloor, 1976; Knorr-Cetina, 1991). The results from empirical testing support the normative hypothesis, and so confirm the argument that citations reflect the payment of intellectual debt (Baldi, 1998). Abramo (2018) recently revisited the relevant conceptualizations, intending to spell out some principles leading to a clear definition of the "impact" from research, and above all, of the appropriate citation-based indicator to measure it. Sociological research provides new insights, particularly concerning the role of "trusted social networks" in gathering and citing information (Thornley et al., 2015). Indeed the literature is rich in studies on the citing behavior of scientists, as seen in a review by Bornmann and Daniel (2008), and a further updated review by Tahamtan and Bornmann (2018), on the theoretical and empirical aspects of the citation process. Wouters (1999) stresses that the sciences present many types of citing culture, such that the publications within the different fields tend to share certain properties of citing: researchers in one field, for example, will tend to cite more publications than those in another (e.g. the

biomedical vs. mathematics fields). “A conceptual core that is mutually shared by every one of them cannot be isolated; the various citing cultures resemble one another, as members of one family do. It is possible, of course, to abstract certain general notions and claim that these constitute the core.” (Wouters, 1999).

Scholars of bibliometrics have been particularly interested in the issues of the geographical dimension of new knowledge creation (publications), i.e. the “internationalization” of research, and the spread of its impact (citations), i.e. domestic versus international spillovers, as well as the relation between these two phenomena. There is in fact an extensive literature applying bibliometric approaches to the study of international research collaboration. This includes descriptive analyses of single countries and country clusters, for China (Niu & Qiu, 2014), India (Shrivats and Bhattacharya, 2014), Italy (Abramo, D’Angelo, & Murgia, 2013), the ASEAN member states (Kumar, Rohani, & Ratnavelu, 2014), the BRICS countries (Finardi & Buratti, 2016), and the OECD countries (Choi, 2012). The US National Science Foundation’s report on Science and Engineering Indicators (NSB, 2018) provides an exhaustive compendium of bibliometric data, also serving to examine the trends in international research.

Taking the normative view, in which citation linkages imply a flow of knowledge from the cited to citing authors, several scholars have relied on publication citations to investigate the international flows of scientific knowledge (Mehta, Rysman, & Simcoe, 2010; Van Leeuwen & Tijssen, 2000). Rabkin, Eisemon, Lafitte-Houssat, and McLean Rathgeber (1979) explored global visibility for four departments (botany, zoology, mathematics, and physics) of the universities of Nairobi (Kenya) and Ibadan (Nigeria). At the level of the single field, Stegmann and Grohmann (2001) measured knowledge “export” in the Dermatology & Venereal Diseases category of the 1996 CD-ROM Journal Citation Reports (JCR), and in seven unlisted dermatology journals. Hassan and Haddawy (2013) mapped knowledge flows from the United States to other countries in the “energy” field over the years 1996-2009. Abramo and D’Angelo (2018) tracked international spillovers of Italian knowledge production, in over 200 fields, by analyzing publication citations. Jaffe, Trajtenberg, and Henderson (1993) compared the location of patent citations with that of the cited patents, to investigate the extent to which knowledge spillovers are geographically localized. Abramo, D’Angelo, and Carloni (2019) conceptualized the diffusion of knowledge between countries in terms of a “balance of knowledge flows” (BKF). The authors were then able to measure the share of domestic versus foreign flows generated by a country’s research system, both by field and as compared to other countries.

Several scholars have verified the correlation between a country’s rate of international research collaboration and the impact of its publications (Bordons, Gomez, Fernandez, Zulueta, & Mendez, 1996, for Spain; Abramo, D’Angelo, & Murgia, 2017, for Italy; Kumar, Rohani, & Ratnavelu, 2014, for ASEAN nations; Kim, 2006, for Germany; Tan, Ujum, Choong, & Ratnavelu, 2015, for Malaysia). This same literature stream also verifies the existence of a “center-periphery” pattern within clusters and pairs of countries (Choi, 2012; Luukkonen, Tijssen, Persson, & Sivertsen, 1993; Schubert & Sooryamoorthy, 2010).

Adams (2013) asserts that internationally coauthored papers are more highly cited because the authors are more likely to be doing excellent research. Gingras and Khelfaoui (2018) have shown the presence of a visibility (citation) advantage for the USA, given the heavy presence of American authors in bibliographic repertories such as

Web of Sciences and Scopus. This fact has a knock-on effect on all those countries that collaborate more intensely with the USA.
Bornmann, Adams, and Leydesdorff (2018), analyzing the research output in the natural sciences of three economically advanced European countries (Germany, Netherlands, UK; years 2004, 2009, 2014), observe that "articles co-authored by researchers from Germany or the Netherlands are less likely to be among the globally most highly-cited articles if they also cite "domestic" research (i.e. research authored by authors from the same country)"; but this observation was not confirmed for the UK.
Continuing these lines of investigation, the present work aims to identify the relationship between the "level of internationalization" (according to percentage of foreign authors in the byline) of a country's scientific production, divided into three classes (totally domestic, prevalently domestic, and prevalently international), and the "level of internationalization" of the citing publications, also divided in three classes (totally domestic, totally international, and mixed). This analysis allows us to answer some interesting questions; for example, concerning a country's totally domestic publications - are these more cited by totally domestic, totally international or mixed publications?
The empirical analysis is based on the Italian publications of the three-year period 2010-2012 and on the citations accumulated up to 31/05/2017. The details of the dataset and methodology are illustrated in the following section, while section 3 presents the results of the analyses. Finally, section 4 discusses the main findings of the work and their implications.

Data and method

Our analysis is based on the Italian-National Citation Report (I-NCR) by Clarivate Analytics, obtained by extracting all publications authored by Italian organizations from the seven main WoS core collection indexes,[i] i.e. publications with at least one affiliation showing "Italy" as country. For the 2010-2012 period, the I-NCR contains a total of 255399 records.[ii] Of these, 17401 show no author-affiliation link, and are therefore excluded from the analysis. The remaining 237998 constitute the dataset; we divide this into three subsets, defined as:
"totally domestic", if all authors in the byline show Italian affiliations only;
"prevalently domestic", if authors affiliated with Italian organizations compose less than 100%, but at least 50% of all coauthors listed in the byline;
"prevalently international": if authors affiliated with Italian organizations compose less than 50% of all coauthors listed in the byline.
To account for multiple affiliations of a single author we adopt a fractional counting method. In case of authors with n affiliations, we assign $1/n$ to each of her/his addresses. This procedure becomes critical only in the case of authors with both domestic and foreign affiliations, in then determining whether the publication is prevalently domestic or international.
The same logic could be applied to classify the citing publications, observed up to 31/05/2017,[iii] however for these the I-NCR does not indicate the author-affiliation link . Being unable to measure the prevalence of authors (Italian or foreign), we thus classify the publications into three subsets:
"totally Italian": if the citing publication has an address list with only Italian addresses;
"totally foreign": if the citing publications has an address list without any Italian address;

"mixed": if the citing publications has an address list with both Italian and foreign addresses.
The following publication exemplifies the assignment to subsets:

Acocella, N., Di Bartolomeo, G., & Pauwels, W. (2010). Is there any scope for corporatism in macroeconomic policies? *Empirica*, 37(4), 403-424.

Author-affiliation list: [Acocella, Nicola] Univ Roma La Sapienza, Dipartimento Studi Geoecon, I-00161 Rome, Italy; [Di Bartolomeo, Giovanni] Univ Teramo, Teramo, Italy; [Pauwels, Wilfried] Univ Antwerp, B-2020 Antwerp, Belgium

The analysis of affiliations reveals that two of the three authors are affiliated with Italian organizations, so the publication is classified as "prevalently domestic".
The work in question has received three citations, as follow.

Acocella, N., & Di Bartolomeo, G. (2013). The cost of social pacts. *Bulletin of Economic Research, 65*(3), 238-255.

Address list: Univ Roma La Sapienza, Rome, Italy; Univ Teramo, I-64100 Coste S Agostino, Italy

Schneider, F., Haigner, S. D., Jenewein, S., & Wakolbinger, F. (2014). Institutions of conflict management and economic growth revisited: A short note. *Empirica, 41*(3), 577-587.

Address list: Univ Linz, Dept Econ, A-4040 Linz, Austria; Univ Innsbruck, Dept Publ Econ, A-6020 Innsbruck, Austria; Gesell Angew Wirtschaftsforsch, Innsbruck, Austria

Colombo, E., Tirelli, P., & Visser, J. (2014). Reinterpreting social pacts: Theory and evidence. *Journal of Comparative Economics, 42*(2), 358-374.

Address list: Univ Milano Bicocca, I-20126 Milan, Italy; Univ Amsterdam, AIAS, NL-1018 TV Amsterdam, Netherlands

The first citation presents an address list indicating only Italy (by the two authors who produced the cited publication), and therefore is classified as "totally Italian"; the second presents only Austrian addresses, and so is classified as "totally foreign"; the third presents an Italian and a Dutch address, and is classified as "mixed".
The following section presents the analyses relating the three categories of cited publications with the three categories of their relative citing ones. After presenting the data at the overall level, we detail the analysis at the level of the macro-scientific domain (disciplinary area). To do this, each publication is first assigned to the subject category (SC) of its hosting journal, conference or book,[iv] as per the WoS classification scheme, and so to the relative macro-disciplinary area.[v]

Results

Out of the dataset of 233998 publications, 138313 (58.1%) are totally domestic, 46433 (19.5%) are prevalently domestic, and 53252 (22.4%) are prevalently international (Table 1). The share of publications cited is directly related to the level of internationalization: the prevalently international ones are cited in 82.3% of cases, prevalently and totally domestic ones in respectively 80.4% and 68.2% of cases. Altogether the publications of the dataset received 3282334 citations, of which almost

three quarters were from totally foreign publications, while the remainder were almost equally from totally Italian (13.1%) or "mixed" (12.5%) ones.
However, this distribution of citing publications differs considerably in relation to the level of internationalization of the cited publication.
The totally domestic Italian publications receive only 67.0% of their citations from totally foreign publications, compared to 70.3% for the prevalently domestic and 82.9% for the prevalently international ones. The correlation between the level of internationalization of the cited and citing publications is also evident also considering the other cells of the matrix: totally domestic products receive 22.7% of their total citations from totally Italian publications, against 12.7% for the prevalently domestic and only 4.9% for the prevalently international ones.
To test the statistical significance of the relationship between the two dimensions in analysis we calculate Pearson's χ^2: the value observed (19.6E+4 with p-value 0.000) unequivocally demonstrates the significance of the association between the level of internationalization of Italian publications and that of the relative citing publications. Cramer's V test, with control for sample size, provides a further measure of the strength of the association and its practical significance: in fact the value (0.173) indicates a rather weak but still significant association between the two variables.

Table 1: Distribution of 2010-2012 Italian publications in function of their level of internationalization and the relative citing publications

Type of Italian publications	No. of Italian publications	Of which cited	Total citations	Of which totally Italian	Of which mixed	Of which totally foreign
Totally domestic	138313 (58.1%)	68.2%	1227947 (37.4%)	22.7%	10.3%	67.0%
Prevalently domestic	46433 (19.5%)	80.4%	670833 (20.4%)	12.7%	17.0%	70.3%
Prevalently international	53252 (22.4%)	82.3%	1383554 (42.2%)	4.9%	12.2%	82.9%
Total	237998	73.7%	3282334	13.1%	12.5%	74.4%

Both intensity of publication (D'Angelo & Abramo, 2015) and citation behavior are known to vary across disciplines. We could then expect that the level of internationalization of scientific production and that of the relevant citing publications would also vary across disciplines. For this reason we repeat the analysis just presented, at the level of single disciplinary area: Tables 2 and 3 report the results.
In Art and Humanities, three publications out of every four are authored by scholars who are all affiliated only with Italian institutions. In Biomedical research, Clinical medicine and Engineering, the incidence of totally domestic publications is higher than 60%; in all the other areas it is higher than 50%, with the sole exception of Physics (43.4%). On the other hand, in Art and Humanities, only 9.4% of publications are prevalently international. Values of less than 20% are also recorded in Mathematics (17.1%) and Engineering (17.7%). Physics again represents the exception, showing a peak of 31.9% of authored publications by prevalently international teams, followed by Earth and space sciences (25.4%) and Chemistry (21.9%).
With regard to the incidence of cited publications out of the total, we observe an increasing trend with the classes of increasing internationalization, meaning in progressing from totally to prevalently domestic to prevalently international. This is

observed in all disciplinary areas except Art and humanities, Biomedical research, Economics, and Law, political and social sciences - areas whose contents are clearly more dependent on the national context. Regardless of the level of internationalization of the Italian publications, the citations come mainly from "totally foreign" citing articles. The highest value is found in Clinical medicine (78.4% at the overall level) but percentages higher than 75% are also observed in other biomedical areas. The "totally Italian" citing papers more often refer to "domestic" papers than "international" ones: in Earth and space sciences the incidence of the totally Italian citing publications drops from 30.8% for totally domestic publications to 4.8% for prevalently international ones. Conversely, the incidence of totally foreign citing publications increases with the level of internationalization of the Italian publications in reference: in Earth and space sciences there is again a difference in incidence of almost 30 percentage points (53.4% for totally domestic vs. 81.8% for prevalently domestic).

The tests of association between the two nominal variables (type of cited and citing publications) are reported in Table 3. The calculations of Pearson χ^2 and corresponding p-value indicate a significant association; the values of Cramer's V, never higher than 0.25, reveal that the association is moderate in Art and Humanities, Earth and space sciences and Physics, but otherwise almost always weak, with the minimum value observed in Biomedical research.

Table 2: Distribution of 2010-2012 Italian publications in function of their level of internationalization and the relative citing publications, by disciplinary area

Disciplinary area	Type of Italian publication	No. of Italian publications	Of which cited (%)	Total citations	Of which totally Italian (%)	Of which mixed (%)	Of which totally foreign (%)
Art and humanities	Totally domestic	2491 (75.1%)	20.4	3331 (55.4%)	35.7	13.2	51.1
	Prevalently domestic	516 (15.5%)	46.7	1447 (24.0%)	16.9	17.5	65.6
	Prevalently international	312 (9.4%)	46.5	1240 (20.6%)	7.6	12.0	80.4
	Total	3319	26.9	6018	25.4	14.0	60.6
Biology	Totally domestic	22044 (57.5%)	77.7	244158 (41.7%)	22.8	10.0	67.2
	Prevalently domestic	8087 (21.1%)	83.7	131362 (22.5%)	13.0	15.0	72.0
	Prevalently international	8192 (21.4%)	87.7	209432 (35.8%)	4.6	9.4	85.9
	Total	38323	81.1	584952	14.1	10.9	75.0
Biomedical research	Totally domestic	27283 (62.2%)	65.3	263432 (39.9%)	19.8	8.7	71.4
	Prevalently domestic	7053 (16.1%)	77.5	124481 (18.9%)	12.8	13.6	73.6
	Prevalently international	9541 (21.7%)	76.2	271704 (41.2%)	6.1	9.0	85.0
	Total	43877	69.7	659617	12.8	9.8	77.4
Chemistry	Totally domestic	9777 (54.9%)	93.2	160606 (46.4%)	22.1	9.5	68.4
	Prevalently domestic	4121 (23.2%)	94.1	84989 (24.5%)	12.6	15.7	71.7
	Prevalently international	3901 (21.9%)	94.5	100649 (29.1%)	4.3	10.2	85.5
	Total	17799	93.7	346244	14.6	11.2	74.2
Clinical medicine	Totally domestic	48525 (63.5%)	63.4	399536 (37.5%)	20.2	8.4	71.4
	Prevalently domestic	11457 (15.0%)	75.4	172282 (16.1%)	12.4	13.8	73.8
	Prevalently international	16402 (21.5%)	77.5	495003 (46.4%)	5.3	9.2	85.6
	Total	76384	68.2	1066821	12.0	9.6	78.4
Earth and space sciences	Totally domestic	7262 (51.3%)	86.5	86476 (37.7%)	30.8	15.8	53.4
	Prevalently domestic	3311 (23.4%)	91.6	52623 (22.9%)	15.9	21.6	62.5
	Prevalently international	3595 (25.4%)	92.5	90294 (39.4%)	4.8	13.4	81.8
	Total	14168	89.2	229393	17.1	16.2	66.7
Economics	Totally domestic	3128 (53.7%)	58.7	15663 (39.3%)	21.5	10.2	68.3
	Prevalently domestic	1504 (25.8%)	73.7	12892 (32.4%)	8.6	12.2	79.1
	Prevalently international	1190 (20.4%)	71.3	11250 (28.3%)	5.4	10.8	83.8
	Total	5822	65.1	39805	12.8	11.0	76.2
Engineering	Totally domestic	29276 (62.2%)	68.7	225875 (49.8%)	26.1	10.3	63.7
	Prevalently domestic	9499 (20.2%)	76.7	104463 (23.0%)	13.6	16.2	70.2
	Prevalently international	8320 (17.7%)	78.0	123441 (27.2%)	5.5	11.1	83.4
	Total	47095	71.9	453779	17.6	11.9	70.5
Law, political and social sciences	Totally domestic	3299 (58.5%)	48.3	12675 (40.4%)	25.8	10.2	64.0
	Prevalently domestic	1213 (21.5%)	66.8	7815 (24.9%)	10.6	14.4	75.0
	Prevalently international	1130 (20.0%)	66.5	10860 (34.6%)	4.3	11.5	84.2
	Total	5642	55.9	31350	14.5	11.7	73.8
Mathematics	Totally domestic	4664 (50.4%)	74.3	27418 (42.2%)	31.2	15.2	53.6
	Prevalently domestic	3011 (32.5%)	82.6	22968 (35.3%)	13.9	23.1	63.0
	Prevalently international	1579 (17.1%)	83.1	14641 (22.5%)	6.9	15.9	77.2
	Total	9254	78.5	65027	19.6	18.1	62.2
Physics	Totally domestic	16069 (43.4%)	75.2	156229 (25.6%)	24.3	14.6	61.1
	Prevalently domestic	9124 (24.6%)	86.1	135337 (22.2%)	11.7	24.4	63.9
	Prevalently international	11827 (31.9%)	89.6	317703 (52.1%)	3.6	20.8	75.6
	Total	37020	82.5	609269	10.7	20.0	69.3
Psychology	Totally domestic	2088 (57.3%)	62.0	15468 (37.9%)	24.2	11.9	63.9
	Prevalently domestic	823 (22.6%)	80.0	10517 (25.8%)	12.9	16.3	70.8
	Prevalently international	733 (20.1%)	86.6	14837 (36.3%)	5.0	10.0	85.0
	Total	3644	71.0	40822	14.3	12.3	73.4

Table 3: Tests* of association between the level of internationalization of Italian publications and that of the relative citing publications, by disciplinary area

Disciplinary area	Obs	Pearson χ^2	p-value	Cramer's V
Art and humanities	3319	0.5E+03	0.000	0.202
Biology	38323	3.5E+04	0.000	0.172
Biomedical research	43877	2.6E+04	0.000	0.139
Chemistry	17799	1.8E+04	0.000	0.162
Clinical medicine	76384	5.1E+04	0.000	0.155
Earth and space sciences	14168	2.4E+04	0.000	0.231
Economics	5822	1.8E+03	0.000	0.152
Engineering	47095	2.7E+04	0.000	0.172
Law, political and social sciences	5642	2.4E+03	0.000	0.194
Mathematics	9254	4.8E+03	0.000	0.192
Physics	37020	5.0E+04	0.000	0.202
Psychology	3644	2.7E+03	0.000	0.180

** (three levels of both the internationalization of the cited publications and the citing publications) => 3x3 contingency matrices = 4 degree of freedom*

Discussion and conclusions

This work analyzes the relation between the level of internationalization of Italian scientific production and that of the relevant citing publications, at an aggregate level and by disciplinary area. The analysis is conducted using bibliometric techniques and as always, the limitations and assumptions embedded in such analyses apply. Caution is therefore recommended in interpreting the findings. The scientific production examined is from the period 2010-2012, with the relevant citing publications observed until 31/05/2017.

The results at aggregate level reveal that more than half of Italian scientific production is totally domestic, while the remaining part is divided almost equally between prevalently domestic and prevalently international. Interestingly, more than half of the citations arrive from totally foreign publications, regardless of the level of internationalization of the cited publications. This finding opens to question the consolidated thesis of the geographical proximity of knowledge spillovers, previously demonstrated through the analysis of patent citations (Jaffe, Trajtenberg, & Henderson, 1993): unlike technical knowledge encoded in patents, scientific knowledge appears to flow rather easily across national borders. Further analysis on the geographical distance of scientific knowledge spillovers has been initiated by the authors.

The factor of the more or less domestic level of a publication clearly has an influence on its interest to foreign scholars, as measured by the level of internationalization of the citing publications. In fact totally domestic publications receive quotations from totally Italian publications to a greater extent than do other categories. At the moment we can only surmise whether this phenomenon is more attributable to a lower average quality of totally domestic publications, or to the more country-specific nature of the research problems addressed. The authors therefore plan further research, to compare the average quality of each publication category along the internationalization dimension.

The in-depth analysis for each disciplinary area confirms the evidence of the overall analysis, with minor differences. For all disciplinary areas, the share of totally domestic publications is above 50% (with the sole exception of Physics), with a peak in Art & Humanities (75%), a discipline which is clearly quite country specific. In Physics, the total domestic publications are only 43%, reflective of a research area tending to require

international collaborations, often involving very large teams, (i.e. high energy physics, and astrophysics).
Cited publications as a share of the total production increase with the level of internationalization, in all disciplinary areas except Art and humanities, Biomedical research, Economics, and Law, political and social sciences. However, examining in more detail, this result emerges from the comparison between totally and prevalently domestic publications, and not from the prevalently domestic/international comparison. With regard to citing publications, in all cases, the analyses at the level of the disciplines confirms that as internationalization of the cited publications increases then so does the incidence of totally foreign citing publications.
An interesting observation is that the totally-Italian citing publications tend more to cite totally domestic publications in the specific areas of Art & Humanities, Earth and space science and Mathematics. The first two of these disciplinary areas are clearly cases where in many subfields the "geographical" context is particularly important. The resulting publications might therefore be of interest to the international scientific community for their methodological content more than for the results themselves, which would be difficult to export or replicate in foreign contexts.
Since knowledge is cumulative, the question arises as to the extent that the findings observed in the Italian context would differ from those of other countries, which will have different of domestic stocks and level of knowledge. In fact, as noted by Abramo, D'Angelo, and Carloni (2019), there is an ever greater chance that new knowledge will stem from domestic research rather than foreign, as: i) the larger is the country in terms of number of researchers and research resources; ii) the more productive is the research system; and iii) the more scientifically advanced the country is in its stock and level of accumulated knowledge, compared to other countries. Given this, it would be interesting to extend this type of analysis to other countries.
A further direction could also be to strengthen the analytical model, particularly in better defining the internationalization of the citing publications, subject to the availability of the author-affiliation link for citing publications. First, the category "mixed" could be further divided between prevalently domestic and prevalently foreign. Second, the totally foreign category could be distinguished in reference to the bylines, showing either single or multiple countries.
Finally, the Italian academic and research systems have recently been subject to reforms, expected to impact the behavior of organizations and their individual scholars. It would therefore be interesting to extend the analysis to periods subsequent to 2010-2012, to assess whether and how the association between the investigated dimensions has evolved. A specific question, for example, could be whether there has been any change in the incidence of totally Italian publications citing totally domestic publications.

References

Abramo, G., D'Angelo, C. A., & Carloni, M. (2019). The balance of knowledge flows. *Journal of Informetrics*, 13(1), 1-9.

Abramo, G. (2018). Revisiting the scientometric conceptualization of impact and its measurement. *Journal of Informetrics*, 12(3), 590-597.

Abramo, G., D'Angelo, C. A. (2018). Who benefits from a country's scientific research? *Journal of Informetrics*, 12(1), 249-258.

Abramo, G., D'Angelo, C. A., & Murgia, G. (2017). The relationship among research productivity, research collaboration, and their determinants. *Journal of Informetrics*, 11(4), 1016-1030.

Abramo, G., D'Angelo, C. A., & Murgia, G. (2013). The collaboration behaviors of scientists in Italy: A field level analysis. *Journal of Informetrics*, 7(2), 442–454.

Adams, J. (2013). Collaborations: The fourth age of research. *Nature*, 497(7451), 557–560.

Baldi, S. (1998). Normative versus social constructivist processes in the allocation of citations: A network-analytic model. *American Sociological Review*, 829-846.

Bloor, D. (1976). *Knowledge and Social Imagery*. London: Routledge, Kegan and Paul.

Bordons, M., Gomez, I., Fernandez, M. T., Zulueta, M. A., & Mendez, A. (1996). Local, domestic and international scientific collaboration in biomedical research. *Scientometrics,* 37(2), 279–295.

Bornmann, L., Adams, J., & Leydesdorff, L. (2018). The negative effects of citing with a national orientation in terms of recognition: National and international citations in natural-sciences papers from Germany, the Netherlands, and the UK. *Journal of Informetrics*, 12(3), 931-949.

Bornmann, L., & Daniel, H. D. (2008). What do citation counts measure? A review of studies on citing behavior. *Journal of Documentation*, 64(1), 45-80.

Choi, S. (2012). Core-periphery, new clusters, or rising stars?: International scientific collaboration among "advanced" countries in the era of globalization. *Scientometrics,* 90(1), 25–41.

D'Angelo, C.A., Abramo, G. (2015). Publication rates in 192 research fields. In A. Salah, Y. Tonta, A.A.A. Salah, C. Sugimoto (Eds) *Proceedings of the 15th International Society of Scientometrics and Informetrics Conference* - (ISSI - 2015) (pp. 909-919). Istanbul: Bogazici University Printhouse.

Finardi, U., & Buratti, A. (2016). Scientific collaboration framework of BRICS countries: an analysis of international coauthorship. *Scientometrics,* 109(1), 433–446.

Gingras, Y., & Khelfaoui, M. (2018). Assessing the effect of the United States' "citation advantage" on other countries' scientific impact as measured in the Web of Science (WoS) database. *Scientometrics,* 114(2), 517–532.

Hassan, S.-U., & Haddawy, P. (2013). Measuring international knowledge flows and scholarly impact of scientific research. *Scientometrics,* 94(1), 163–179.

Jaffe A.B., Trajtenberg M., Henderson R. (1993). Geographic localization of knowledge spillovers as evidenced by patent citations. *The Quarterly Journal of Economics*, 108 (3), 577-598.

Kaplan, N., (1965). The norms of citation behavior: Prolegomena to the footnote. *American Documentation*, 16(3), 179–184.

Kim, K. W. (2006). Measuring international research collaboration of peripheral countries: taking the context into consideration. *Scientometrics,* 66(2), 231–240.

Knorr-Cetina, K. (1991). Merton sociology of science: the first and the last sociology of science. *Contemporary Sociology-a Journal of Reviews*, 20, 522-526.

Kumar, S., Rohani, V. A., & Ratnavelu, K. (2014). International research collaborations of ASEAN Nations in economics, 1979–2010. *Scientometrics,* 101(1), 847–867.

Luukkonen, T., Tijssen, R. J. W., Persson, O., & Sivertsen, G. (1993). The measurement of international scientific collaboration. *Scientometrics,* 28(1), 15–36.

Martin, B. R., & Irvine, J. (1983). Assessing basic research: some partial indicators of scientific progress in radio astronomy. *Research Policy*, 12(2), 61-90.

Mehta, A., Rysman, M., & Simcoe, T. (2010). Identifying the age profile of patent citations. *Social Science Research Network*, 25(7), 1179–1204.

Merton, R. K. (Ed.) (1973). *The sociology of science: theoretical and empirical investigations*, Chicago, USA, University of Chicago Press.

Mulkay, M. (1976). Norms and ideology in science. *Social Science Information*, 15 (4-5), 637-656.

National Science Board-NSB (2018). *Science and Engineering Indicators 2018*. NSB-2018-1. Alexandria, VA: National Science Foundation.

Niu, F., & Qiu, J. (2014). Network structure, distribution and the growth of Chinese international research collaboration. *Scientometrics,* 98(2), 1221–1233.

Rabkin, Y. M., Eisemon, T. O., Lafitte-Houssat, J.-J., & McLean Rathgeber, E. (1979). Citation visibility of Africa's science. *Social Studies of Science*, 9(4), 499–506.

Schubert, T., & Sooryamoorthy, R. (2010). Can the centre-periphery model explain patterns of international scientific collaboration among threshold and industrialised countries? The case of South Africa and Germany. *Scientometrics,* 83(1), 181–203.

Shrivats, S. V., & Bhattacharya, S. (2014). Forecasting the trend of international scientific collaboration. *Scientometrics,* 101(3), 1941-1954.

Stegmann, J., & Grohmann, G. (2001). Citation rates, knowledge export and international visibility of dermatology journals listed and not listed in the Journal Citation Reports. *Scientometrics,* 50(3), 483–502.

Tahamtan, I., & Bornmann, L. (2018). Core elements in the process of citing publications: Conceptual overview of the literature. *Journal of Informetrics*, 12(1), 203-216.

Tan, H. X., Ujum, E. A., Choong, K. F., & Ratnavelu, K. (2015). Impact analysis of domestic and international research collaborations: A Malaysian case study. *Scientometrics,* 102(1), 885–904.

Thornley, C., Watkinson, A., Nicholas, D., Volentine, R., Jamali, H. R., Herman, E., ... & Tenopir, C. (2015). The role of trust and authority in the citation behaviour of researchers. *Information Research*, 20(3), 1-17.

Van Leeuwen, T., & Tijssen, R. (2000). Interdisciplinary dynamics of modern science: Analysis of cross-disciplinary citation flows. *Research Evaluation*, 9(3), 183–187.

Wouters, P. F. (1999). *The citation culture* (Doctoral dissertation, Universiteit van Amsterdam).

---

[i] SCI-E: Science Citation Index Expanded; SSCI: Social Sciences Citation Index; A&HCI: Arts & Humanities Citation Index; CPCI-S: Conference Proceedings Citation Index- Science; CPCI-SSH: Conference Proceedings Citation Index- Social Science & Humanities; BKCI-S: Book Citation Index– Science; BKCI-SSH: Book Citation Index– Social Sciences & Humanities.

[ii] With no restrictions to any document type.

[iii] We include self-citing publications, because it should not matter whether the development flowing from a publication is performed by the same author(s), as long as it is performed in her or his country.

[iv] Publications hosted in multi-category sources are assigned to each of the SCs.

[v] Our assignment of SCs to disciplinary areas (Mathematics; Physics; Chemistry; Earth and Space Sciences; Biology; Biomedical Research; Psychology; Clinical Medicine; Engineering; Economics; Law, political and social sciences) follows a pattern previously published in the ISI Journal Citation Reports website, although this information is no longer available through the Clarivate web portal.

Comparing institutional-level bibliometric indicator values based on different affiliation disambiguation systems. Benchmarking Web of Science and Scopus platform tools against a gold-standard data set for Germany

Paul Donner[1], Christine Rimmert[2], and Nees Jan van Eck[3]

[1] *donner@dzhw.eu*
German Centre for Higher Education Research and Science
Schützenstr. 6a, 10116 Berlin, Germany

[2] *ecknjpvan@cwts.leidenuniv.nl*
Centre for Science and Technology Studies, Leiden University
P.O. Box 905, 2300 AX Leiden, The Netherlands

[3] *christinerimmert@mailbox.org*
Institute for Interdisciplinary Studies of Science, Bielefeld University
P.O. Box 10 01 31, 33501 Bielefeld, Germany

Abstract

The present study is an evaluation of three bibliometrically pertinent institution disambiguation systems. The Web of Science normalized institution name and Organization Enhanced system and the Scopus Affiliation ID are tested against complete, independent institution disambiguation system for a sample of German public sector research organizations, used as a gold standard. We study in particular the accuracy of a number of common bibliometric indicators of institutions under the different disambiguation systems. This is the first published study of this kind. The main finding is that none of the studied systems can be used to obtain accurate bibliometric indicator values for the sample institutions.

Introduction

Scientometric studies at the level of research institutions face the challenge of the correct attribution of publications to institutions by means of disambiguating the heterogeneous address data in the author-provided affiliation information present in publications and recorded in bibliographic databases. At present, institutional affiliation information in academic publications is not standardized and unique identifiers for research institutions have not yet been adopted. Therefore, in order to generate valid primary data on publications for studies at the meso-level, the assignment of address strings to known real institutional entities is a necessary and crucial procedure. Institution name disambiguation belongs to a class of problems known as named entity normalization, in which variant forms have to be matched to the correct preferred form. Another prominent member of this class is author name disambiguation. Disambiguated affiliation information can contribute to the performance of author name disambiguation systems which employ affiliations as background information. In the recent past, a nearly complete institutional disambiguation for German research institutions was developed and implemented at the Institute for Interdisciplinary Studies of Science of Bielefeld University as a major component of a national bibliometric data infrastructure for research and monitoring. The system has been tested and improved over a number of years and is now in production use. This allows us to study the degree to which the use of a sophisticated disambiguation system, with near-complete national scale coverage, actually leads to different bibliometric indicator values compared to a situation when no such system is available and simpler alternatives to the attribution problem have to be used. We consider here (1) the case where a very simple unification strategy using *ad hoc* lexical searches in address data fields in

a bibliographic database is conducted in order to collect publications of the target institutions (based on vendor pre-processed affiliation data in Web of Science) and (2) the use of bibliographic database vendors' own institution disambiguation systems (in both Web of Science and Scopus). We believe these two situations are common in practice outside of specialized research or evaluation units with access to the raw data of bibliographic databases. The performance and implications of these approaches is therefore relevant and of wide interest.

Systems, data, and methods

Disambiguation system for German research institutions

In this section we only give a summary of the disambiguation system that was developed for disambiguating German institutions. For a full description of the system, the readers are referred to Rimmert *et al.* (2017). The system, which we refer to as the KB system, is comprised of a set of known and uniquely identified German research institutions, a mapping of institutions to affiliation records identified as belonging to each institution for the two data sources WoS and Scopus, a hierarchical classification of institutions into sectors, and furthermore, a change history of institutions which records the splitting and merging and incorporation of institutions and sector changes. This tracking of structural changes affords the application of different handling of structural changes due to different project contexts. In the KB system two different analytical views are implemented. With Mode S (for synchronic allocation) we can perform analyses that take into account the institutional structure as it was at the time of publication for each paper. Institutions which have later come to be related to another institution through structural changes such as e.g. mergers or splits are treated as different entities. In Mode A (asynchronic, current perspective) on the other hand, we can analyze publication records of institutions at present, including publications of predecessor units. The mapping of institutions to affiliation records is a deterministic rule-based classification. The core of the institutional coding procedure is a mapping of author addresses to the corresponding uniquely identified research institutions and their subdivisions, using a large library of regular expressions which currently contains some 45,000 expressions and is continuously expanded and improved.

The sector classification contains the classes of higher education sector (universities and universities of applied sciences), four major non-university research organizations (Fraunhofer-Gesellschaft (FHG), Helmholtz Association (HGF), Leibniz Association (WGL) and Max Planck Society (MPG)), private companies, registered associations, government laboratories, and academies of science. For the sector information, structural changes over time and multiple assignments of research institutions to sectors are also available.

The version of the KB system which was used for this study contained 2,097 institutions, which also includes placeholder records for unidentified records. An evaluation of the KB disambiguation system was conducted prior to the main study. We omit this evaluation due to space constraints, but the results can be requested from the first author. We conclude based on the good results of this evaluation that the KB system is a valid gold standard benchmark for German institutional affiliation disambiguation data.

Disambiguation systems of Web of Science and Scopus

We deliberately do not attempt to describe the workings of the proprietary disambiguation systems of WoS and Scopus and regard them as black boxes, of which we only analyze the results. The reason for this is that both systems are not documented in any detail by the providers. What we can gather from the information of the platforms is that WoS Organizations Enhanced (OE) is based on lists of variant names mapped to preferred names (https://clarivate.libguides.com/woscc/institution accessed 04/13/2018). Regarding the Scopus

Affiliation Identifiers (AFIDs), the documentation merely informs us that "The Affiliation Identifier distinguishes between affiliations by assigning each affiliation in Scopus a unique number and grouping together all of the documents affiliated with an organization" https://service.elsevier.com/app/answers/detail/a_id/11215/supporthub/scopus/related/1/ accessed 04/13/2018. No information is given about how the system works.

Data

The data used in the analyses is derived from the licensed source data of Web of Science (Science Citation Index Expanded, the Social Sciences Citation Index, the Arts & Humanities Citation Index, and the Conference Proceedings Citation Index) and Scopus obtained in spring 2017. The data were loaded into in-house relational databases, cleaned, and enhanced. The most important enhancement is the disambiguation of German author addresses to known German research institutions, conducted separately for each data source using the KB disambiguation system described in the previous subsection. The units of analysis are academic institutions, in particular German universities, universities of applied sciences and non-university research institutes. Publications are restricted to articles and reviews, published between 1995 and 2017. To be included, an institution had to have at least 50 such publications associated with it according to the KB disambiguation of the WoS data. This resulted in 445 institutions. This list of institutions is used for WoS and Scopus data based analyses. Notice that we do not attempt to use an identical set of publications for the two data sources.

Scopus AFID

For the Scopus data we compare the reference data to sets of publications that have one or more common assigned AFIDs (affiliation identifiers) as provided by Elsevier. Some preprocessing steps to align the Scopus and KB disambiguation systems were performed in order to make them comparable, as they are conceptually and structurally somewhat different. To match AFIDs to KB system IDs, the AFID(s) for each institution in our sample was obtained by searching the Scopus online platform. It is not clear if and how the definition of an institution in Scopus differs from the one that the disambiguation is based on. One difference that we have taken into account is that the AFID system typically has separate IDs for university hospitals and the universities they belong to, which is not the case in the KB system. We have therefore merged those AFIDs to create more comparable and consistent publication record sets. Furthermore, in some cases, more than one AFID for the same institution exists in Scopus, for example for multiple branch locations. If these are logically linked in the hierarchical relations in the Scopus system, we also merged these linked AFIDs. If not, we have only taken the most commonly used AFID per institution. We found that in the AFID system, publications with affiliations referring to predecessor units are grouped with their current unit. Based on this observation we compare the AFID results with those from the KB system's A mode.

WoS Organizations Enhanced

The WoS OE system does not have unit identifiers, but Preferred Names, which are additionally assigned as institution names to affiliations considered to belong to one real institution. In order to identify the WoS Preferred Name for the institutions in our set we started by identifying all Preferred Names of records with German addresses which occur more than 20 times. From this list we chose the Preferred Name matching the target institution name, otherwise excluded the institution from this part of the study. In fact, for our sample set, it is not possible to retrieve the corresponding publications on main institutional level in the majority of cases. While many universities are recorded in Organization Enhanced, institutions of Fraunhofer Gesellschaft (FHG), Helmholtz Gemeinschaft (HGF), Leibniz Gemeinschaft (WGL), and Max-Planck-Gesellschaft (MPG) are almost all grouped such that only all publications of FHG, HGF, WGL

and MPG, respectively, can be found but rarely their member institutes. Similar to AFID, also in in the WoS OE system, predecessor units are grouped under the preferred name of the current institution. In consequence, we also compare the WoS OE system with mode A of the KB system.

WoS institution name search

Besides the comparison of OE data with the KB disambiguation we also investigate the performance of a lexical search by institution name in the WoS affiliation data. As pointed out directly above, the coverage of institutions in the OE systems is very poor, which supports the notion that this approach is sensible. The institution name search method does in part make use of WoS disambiguation efforts, because institution names extracted from affiliation information in papers are not indexed identical to how they are given in the original publication, but are normalized. Because the affiliation in Scopus is not transformed, we do not apply a similar search strategy to Scopus data. In fact, it is not possible to conduct comparable searches across these two databases due to fact that WoS only contains these normalized address strings while Scopus contains the original address strings. In this scenario we model a hypothetical user who has a list of the names of the German research institutions available which is used as a basis for generating search terms. We also assume that the user is familiar with searching in WoS data to a sufficient degree. This scenario further requires a definition of the name list, the search terms and the search parameters.

In order to generate a plausible name list, we begin by using the KB institutional disambiguation results to find the most common normalized short name in the WoS data for each real institution in our initial set. We then manually assessed the lists side-by-side with the real names and discarded any short name that cannot be deduced from the name list, using instead the next most common name variant iteratively. This relates to our decision of going beyond a completely naïve and automatic procedure and including a realistic degree of user common sense and domain familiarity. We use the search term list as retrieval inputs, while also ignoring capitalization and allowing truncation at the end of the term and search the full address information field with them. This comes reasonably close to an informed, but nonspecialized, search for an institution on the online platform of WoS. It is general in the sense that all institutions are treated in the same way and no special knowledge of affiliation idiosyncrasies is included. It is however limited in that we only consider one name variant per institution.

Because we directly use the normalized affiliation data as it is indexed in WoS, it is clear that we use the normalized versions of the institution names at the time of publication. Thus we use the S mode of the KB system for comparison.

Evaluation design and indicators

To assess performance of the studied systems in terms of being able to identify the correct publications of the units we use the information retrieval measures of precision and recall. For this task, precision is calculated as the share of correctly retrieved publications among the total number of retrieved publications. Recall is the share of correctly retrieved publications among all retrieved publications. The correct publications are those identified by the KB system.

In order to quantify the effect of the application of a specific institutional disambiguation on scores of bibliometric indicators, we calculated the indicator values for the sets of publications of each institution as retrieved by the KB system, which is again considered as a validated gold standard for the selected institutions in this study, and for each of the three alternative systems. The differences of indicator values are calculated and the arising error distributions are described.

A number of bibliometric indicators were chosen for which the score differences of the disambiguation systems will be quantified. We consider the three domains of publication

output, collaboration, and citation impact. For the latter two domains we have selected both indicators that are absolute numbers and indicators which are ratios or averages, i.e. relative indicators. The indicators are summarized in Table 1. It is clear that the absolute indicator values are dependent on the number of correctly identified publications. It might however be anticipated that the values of relative indicators are less influenced when only part of the correct publication set is used as their input.

Table 1: Overview of selected bibliometric indicators

Domain	*Indicator*	*Remarks*
Publication output	P: number of publications (full count)	The number of publications of an institution
Collaboration	P(collab): number of collaborative publications	The number of an institution's publications that have been co-authored with one or more other institutions
	PP(collab): proportion of collaborative publications	The proportion of an institution's publications that have been co-authored with one or more other institutions
	P(int collab): number of international collaborative publications	The number of an institution's publications that have been co-authored by two or more countries
	PP(int collab): proportion of international collaborative publications	The proportion of an institution's publications that have been co-authored by two or more countries
Citation impact	TCS: total citation score	The total number of citations of the publications of an institution, 5-year citation counts
	MCS: mean citation score	The average number of citations of the publications of an institute, 5-year citation counts
	TNCS: total normalized citation score	The total number of citations of the publicationsof an institution, normalized for field (WoS: Subject Category; Scopus: ASJC) and publication year, 5-year citation counts
	MNCS: mean normalized citation score	The average number of citations of the publications of an institution, normalized for field (WoS: Subject Category; Scopus: ASJC) and publication year, 5-year citation counts
	P(top 10%): number of highly cited publications	The number of an institution's publications that, compared with other publications in the same field and the same year, belong to the top 10% most frequently cited, 5-year citation counts
	PP(top 10%): share of highly cited publications	The proportion of an institution's publications that, compared with other publications in the same field and the same year, belong to the top 10% most frequently cited

Accuracy measures

We compare two vendor-provided disambiguation system results and one search-based result with the KB system's results, which we take as the correct result providing reference values.
We divide the system evaluation into two parts. On the one hand, for each institution in the evaluation set, we would like to find all its publications, without retrieving any publications it was not involved with. This is a basic information retrieval task which can be measured with precision and recall. We also use retrieval performance, including the absolute number of retrieved institutions in the evaluation set to analyze the coverage of the systems with respect to our sample of 445 institutions.
The second component of the evaluation concerns the bibliometric indicator scores calculated from the retrieved institution publication sets. In general, the numerical discrepancy between the indicator values using the KB disambiguation (reference values) and the other methods will be expressed as relative deviation in percent, calculated as:
deviation = (*observed system score* − *KB system reference score*) / *KB system reference score* × 100.

The deviation has a lower bound at −100 % and is unbounded in the positive direction.
For example, let the reference MCS of a unit be 5.5 (calculated based on the KB disambiguated data) and the focal value obtained from a simple institution search in WoS be 4.2. Then the deviation as defined above is (4.2 − 5.5) / 5.5 × 100 = −23.6%. In this case the correct result would be underestimated by 23.6 percent. For each indicator the computed deviations for each institution are collected. Our main measure of accuracy is the percentage of values within a range of ± 5 % of the reference score.

Results

An overview of the coverage of German institutions in the WoS and Scopus systems and achieved by a lexical search in WoS is provided in Table 2.

Table 1. Coverage of Sample Institutions by the Systems

Sector	*Number of Institutions*	*Covered in WoS OE*	*Covered in WoS Search*	*Covered in Scopus AFID*
Fraunhofer-Gesellschaft	62	1	62	57
Helmholtz Association	23	9	23	16
Max Planck Society	86	0	86	78
Universities	107	66	107	96
Leibniz Association	83	8	83	56
Universities of applied sciences	90	9	90	75
Total	445	91	445	376

Note. The figure in the row "Total" may differ from the sum of the above cells because some institutions are assigned to more than one sector

We are only able to find 91 of our 445 (20%) evaluation sample institutions in the OE system. The coverage of OE is the lowest among the systems considered. The covered institution set is comprised mostly of institutions of higher education and does not cover most institutes of the four extra-university research associations as institutions. Using the search strategy, we can find one normalized form for each institution, achieving complete coverage. The Scopus AFID system covers 376 (85%) of the institutions with no conspicuous differences between sectors.

WoS Organizations Enhanced

Note that all results should be interpreted with due caution, as the WoS OE system covers only 91 selected institutions. The precision of WoS OE for these institution publication sets is 0.95 on average across institutions, weighted by publication numbers. Hence, typically about 5% of the returned publications in a result set of a specific Preferred Name will be false positives. The weighted mean of recall across institutions is 0.93, meaning the result sets do not include about 7% of relevant publications on average. The contrast between unweighted (0.87) and weighted mean for recall shows that the results for larger institutions (in terms of number of publication) are better than those for smaller institutions.

We now turn to the results of the comparison of the scores of bibliometric indicators between the WoS OE and the KB system. The results are presented in Table 3 in the form of summaries of the deviation score distributions. It can be seen that absolute indicator scores (number of publications, collaborative papers, and citations) are less often within the range of nearly correct values than relative indicator scores.

Table 3: Deviation of indicator scores of WoS OE from KB system (Mode A), N=91

Indicator	*Percent Deviation Within ±5%*	*Median Absolute Deviation [%]*
P	40.7	4.5
P(collab)	38.5	4.3
PP(collab)	93.4	0.6
P(int collab)	37.4	3.4
PP(int collab)	86.8	1.3
TCS	52.6	4.0
MCS	79.1	1.3
TNCS	40.7	3.7
MNCS	85.7	1.3
P(top 10%)	48.4	3.7
PP(top 10%)	100.0	0.04

WoS institution name search

In this section, we compare the results of the WoS institution name search with those of the KB system. Note that the search makes use of the institution name normalization of WoS and that

we have deliberately searched for the single most common WoS normalized name per institution as described above. Using this search method, we obtain vastly more institution publication sets than using WoS OE, in fact full coverage of all sample institutions. We obtain rather poor results for average precision of 0.61 when weighting institutions by the number of publications and 0.67 for arithmetic mean. Publication sets for this method hence will often contain many publications not actually having authors from the institutions in question. Recall is at 0.74 weighted mean, 0.55 arithmetic mean, which means the publication lists returned by these queries will commonly be very incomplete, but less so for the larger institutions. These results suggest that the normalization procedure of WoS is often not able to group most of the relevant institution name variants under one normalized form. The results of the comparison of bibliometric indicator scores between the WoS institution name search approach and the KB system for Mode S are given in Table 4. The shares of institutions for which the scores obtained with the institution name search approach are within ±5 % of the reference score are low, especially for the absolute indicators. Dispersion of the deviations is very high. Here as well, ratio- and mean-based citation scores are relatively less inaccurate. Evidently, the incomplete publication list result sets of this method lead to substantially inaccurate scores for all indicators.

Table 4: Deviation of indicator scores between WoS institution name search and KB system (Mode S), N=445

Indicator	*Percent deviation within ±5%*	*Median absolute deviation [%]*
P	6.7	52.5
P(collab)	7.2	45.4
PP(collab)	29.0	12.5
P(int collab)	10.1	45.2
PP(int collab)	20.5	17.9
TCS	10.8	39.6
MCS	9.9	18.8
TNCS	5.7	55.5
MNCS	52.0	6.6
P(top 10%)	7.6	46.6
PP(top 10%)	13.7	14.7

Scopus AFID

Precision is quite high 0.96 weighted mean but, in contrast, recall is more moderate at 0.86 weighted mean. Again, we find that the weighted mean precision and recall are slightly greater than the unweighted ones, suggesting that disambiguation quality is typically a little better for larger institutions. We also note that the coverage of our selected benchmarking institutions for the AFID system is 376 out of 445, i.e. 85%. Unlike the WoS OE system, the Scopus AFID system is not largely concentrated on universities. The direct comparison of the results for indicator scores calculated with the Scopus platform disambiguation system, AFID, on the one hand and those calculated with the KB system on the other in terms of distributions of percent

deviation are given in Table 5. We find on average, for the absolute indicators, considerable shares of scores that are outside the range of accepted values. Relative indicators scores are less severely affected, but not within the accepted range often enough to be considered reliable. It is worth pointing out that in particular TCS, the total number of citations, is rarely within the allowed range, which however did not seem to overly affect the other citation indicators.

Table 5: Deviation of indicator scores between Scopus AFID and KB system (Mode A), N=376

Indicator	*Percent deviation within ±5%*	*Median absolute deviation [%]*
P	40.7	9.7
P(collab)	40.7	9.3
PP(collab)	86.7	1,4
P(int collab)	40.8	8.9
PP(int collab)	71.5	2.7
TCS	12.8	15.0
MCS	59.0	4.9
TNCS	41.0	9.7
MNCS	66.5	3.1
P(top 10%)	40.2	8.6
PP(top 10%)	68.1	3.1

Discussion

We have investigated the accuracy of bibliometric indicator values for German publicly funded research organizations that can be obtained by a search strategy on vendor-normalized data (for Web of Science) and the use of database vendors' proprietary disambiguation systems (for both Web of Science and Scopus) in comparison with results from a nearly complete and independent institutional disambiguation with detailed performance characteristics.

We found during our study that conceptual differences between the three disambiguation systems and a lack of documentation of both the WoS and Scopus systems were obstacles to straightforward comparisons. Especially the definition of the basic institutional entity – which is a crucial point for comparing disambiguation systems – varies among the disambiguation systems. In Scopus, e.g., university hospitals are kept separated from university entities. They have different AFIDs which are not connected in any way. This inhibits evaluations for universities including their university hospitals/medical faculties. For a comparison with the KB system, these entities (university hospital and university) had to be aggregated manually. A further issue was the handling of precursor institutions. In order to obtain valid results we evaluated the systems on their own terms, modifying the KB system as necessary which was possible because of its flexibility. In WoS OE, the definition of the institutional entity (e.g. MPG as one single institutional entity) largely rules out a comparison on the institutional level as defined in the KB system for some KB sectors. Furthermore, there is no clear documentation on the handling of structural changes over time such as splits or mergers. For analyses on the institutional level, these issues may lead to severe negative consequences. We find that WoS

Organizations Enhanced has the smallest coverage of our institutions sample at 20 % and is mainly constrained to universities. The coverage of Scopus AFID, on the other hand, is 85 % and not largely limited to one institution type. We were able to obtain complete coverage with a basic search in the normalized affiliation string data for WoS. These results show that the utility of the WoS and Scopus institution disambiguation systems for bibliometrics is very limited, as they do not currently (data as of 2017) provide wide coverage of disambiguated research organizations. In the WoS OE and Scopus AFID systems, precision of the obtainable result sets were close to adequate levels. In both of these systems, recall rates were not sufficiently high to obtain accurate indicator scores. Furthermore, we find substantial variation in precision/recall across institutions, indicating that within one system, these values are not systematically similar across the covered institutions but differ on a case-by-case basis. As for the tested Name Search method on normalized WoS data, precision and recall scores are poor.
Our results indicate that indicator values will typically not be within tolerable error margins on the organizational level, which we have set at ±5 % of the reference value. This holds for plain count indicators and relative indicators. This indicates that precisions/recall values of about 0.95/0.85 are not sufficient to obtain accurate bibliometric indicator values on the institutional level.
Concerning the classes of indicators used in this study, we note that count variable indicator values are directly affected by recall rates and are not usefully reliable under the studied disambiguation systems. Ratio- and average-based indicators, given large enough sample sizes, are more reliable given result set incompleteness. However, our results discourage their use insofar as they are obtained through the studied disambiguation systems, as they are not nearly accurate enough for bibliometric applications. For any use with policy implications the necessary data cleaning procedures for disambiguating affiliation data must be carried out independently for the time being. Relying on vendor disambiguation systems will incur serious inaccuracies in indicator values on the institutional level. We stress that any such study as this one only presents the current situation and that disambiguation systems are improved occasionally. For both vendors' systems, no adequate documentation including performance figures are available, further impeding their adoption in bibliometric studies.

Acknowledgments

This project was partially funded by grant 01PQ17001 from German Federal Ministry for Education and Research.

References

Rimmert, C., Schwechheimer, H., Winterhager, M., (2017). Disambiguation of author addresses in bibliometric databases – technical report. Universität Bielefeld, Institute for Interdisciplinary Studies of Science (I²SoS). https://pub.uni-bielefeld.de/publication/2914944

Measurement of research capacity using disciplinary agglomeration indicators: National university "rankings" in Japan

Masashi Shirabe[1]

[1] *shirabe.m.aa@m.titech.ac.jp*
Tokyo Institute of Technology, Oookayama 2-12-1 W9-77, Meguro, Tokyo 152-8550 (Japan)

Abstract
The author presents a preliminary report on indicators based on a dataset from the biggest and most important academic research funding system in Japan, i.e., the Grants-in-Aid for Scientific Research. Indicators made from ten years (FY 2006-15) worth of data categorized by discipline are analysed. To understand the characteristics of the indicators, they are compared with bibliometricc indicators. The interim results suggest that funding-based indicators are useful for measuring the research capacity of universities while respecting the diversity of their research activities. A preliminary analysis of correlations to bibliometric indicators, however, suggests that the significance of the indicators must be considered carefully, because they are so strongly correlated to the number of papers.

Introduction

Since the ruling Liberal Democratic Party (LDP) published its first "Japan Revitalization Strategy" in 2013, world university rankings have become one of the main concerns for higher education policymakers in Japan. The document sets a target of ten or more Japanese universities being ranked in the top 100 universities in the world within the next decade (but without identifying which rankings), and this target has been incorporated into relevant plans, programs, policies of Japan. Moreover, in the fifth science and technology basic plan (FY2016-20), the government introduced KPIs (key performance indicators) for the first time and was about to determine world university rankings as one of them. In the end, the phrase "international comparisons of universities" replaced "world university rankings," despite that the meaning of "international comparisons of universities" is unclear.
Under these circumstances, one of the top research universities in Japan, Nagoya University, published its ten-year plan, where it described its "Developing Joint Degree Programs with Top University" strategy as follows.

> To join the upper rankings of the Top 100 universities in the world by 2020, Nagoya University seeks to first, increase our faculty members' peer-reviewed publications with international co-authors and second, to raise the percentage of international faculty on our campus. ...
>
> https://www.jsps.go.jp/j-sgu/data/torikumigaiyou/h26-27/sgu_h26-27initiatives_e_a07.pdf

University management as such is like putting the cart before the horse. Moreover, the continuing underachievement of Japanese universities（e.g., Phillips, 2017） has exacerbated the situation.
In this context, we started a project to develop indicators to measure the research capacity of Japanese universities with support from MEXT (the Ministry of Education) in 2016. Our research project is not so much academic as practical and policy oriented. Here, I will make a preliminary report on the indicators that we developed on the basis of a dataset from the biggest and most important academic research funding system in Japan, i.e., the Grants-in-Aid for Scientific Research (https://www.jsps.go.jp/english/e-grants/).

Our project's policy

Although we did not draw a direct line to the "Leiden Manifesto" (Hicks et al., 2015), our project's policy can be considered to have some relation with it. That is, we have developed indicators of the research capacities of universities in line with the following policy.

1. To use bibliometric data and funding (Grants-in-Aid for Scientific Research) data faculty members and university administrators/staff can access
2. To use various indicators that can be applied to a variety of disciplinary fields
3. Not to produce mere rankings of universities, but instead to respect the diversity of universities in their local contexts
4. To provide explanations, limitations, and problems of the indicators and methods

We placed the most importance on 1) measurement of research capabilities *by using research categories that are detailed enough and accepted in academia*, and 2) measurement of research capability not only with activity indicators (e.g., number of papers) and impact indicators (e.g., normalized citation counts) but also *with agglomeration ("ATSUMI" in Japanese) indicators* (e.g., institutional h-index).

Regarding the first requirement, we chose the categorization used in the Grants-in-Aid for Scientific Research system. Actually, this categorization is used by various public research funding organizations and covers a variety of situations; for example, it used as a tag system to describe researchers' expertise. Thus, we believe that this categorization is the best one for our research purposes. However, although this categorization can be used in the analysis of funding data, we cannot apply it to the analysis of bibliometric data. This is because the subject categories of the Grants-in-Aid system are different from those of popular bibliometric databases (e.g., Scopus). Thus, we must re-categorize the journal articles in Scopus at the journal level to enable a full comparison of the bibliometric data and funding data. Despite this drawback, some of the names of the categories in the Grants-in-Aid system are identical to those of Scopus. Thus, two kinds of data can be compared in such categories.

Another point to be considered in the categorization is that for historical reasons, the University of Tokyo has consistently been at the top of pack in terms of research output/input resources and in student recruitment. All other universities are ranked in a hierarchy with the University of Tokyo at the top. This structural factor could easily conceal the unique characteristics of each university if we measured the research capacities of universities by using coarse categories. Therefore, a sufficiently detailed categorization is required; we selected a discipline-level categorization of the Grants-in-Aid, which is the second most detailed categorization in the Grants-in-Aid system.

The second requirement is related to the specific conditions of Japanese universities, especially, those of national universities. That is, if a university has a school (e.g., school of science), the school is likely to consist of a full set of departments (e.g., departments of mathematics, physics, chemistry, and so on). Consequently, the size of each department tends to be small. For example, the department of sociology at the University of Tokyo consists of only 6 academic staffs (http://www.l.u-tokyo.ac.jp/sociology/english/faculty.html), whereas the department of sociology at Cambridge University consists of 19 academic staffs(https://www.sociology.cam.ac.uk/people/academic-teaching-staff-profs). Meanwhile, private universities sometimes show different behavior. For instance, Rikkyo University's department of sociology with 14 academic staffs (http://socio.rikkyo.ac.jp/society/staff/), and the department is part of the college of sociology. Japanese universities vary tremendously in degree of agglomeration of research activities even within a traditional and popular research category. This is also the case in research fields of natural sciences and engineering. Thus, we decided to introduce disciplinary agglomeration indicators, which are not indicators to reflect

the mere size of research activities but to reflect size of research activities over a certain level of "quality."

Furthermore, the fact that functional differentiation of national universities has become a focus of Japanese higher education policies is also strongly related to our project's policy. Since FY 2016, national universities in Japan have been exclusively categorized into just three groups (i.e., "excellent education and research", "excellent education and research in specific fields", and "regional contribution"). More than half of national universities, even those with strengths in specific disciplines, are assigned to the "regional contribution" category. The government started to concentrate non-competitive funding in national universities according to this categorization. The government allocates non-competitive funding to universities with little or no attention to the internal diversity of the universities, so a disciplinary agglomeration in a small field, which may have formed gradually, may end up being dissolved in the near future. To avoid such a consequence, it is important to visualize disciplinary agglomerations by university.

Methods and data

Ten years' (FY 2006-15) worth of data on Grants-in-Aid for Scientific Research will be analysed in this research. Although the same data are downloadable from an open-access database of the funding program, we used the data provided by the funding body (the Japan Society for the Promotion of Science, JSPS).

As mentioned above, we chose the second most detailed categorization (the discipline level). However, as there are several minor changes in the categorization for this period, by considering the most detailed categorization (research field level categorization) in the grant system, funding data were labelled on the basis of the FY2015 discipline level categorization. As a few research fields have to be allocated to two disciplines, we divided up the number of proposals granted and amount of money awarded equally between the two disciplines.

There is a constraint on the number of proposals a researcher can submit in a year as a research representative (i.e., two proposals per year at most). Thus, even if there is a superstar researcher in a discipline in a university, the number of granted proposals in that discipline in that university is limited, and it is rather difficult for the university to earn a lot of research funding in the discipline over the course of ten years. Therefore, the number of proposals granted and amount of money rewarded in a discipline in a university can be indicators of agglomeration of their research activities.

In the data aggregation, the number and the amount were only allocated to the university where a research representative is affiliated. It is theoretically possible to allocate them to the universities where the representative and its collaborator are affiliated; however, we did not adopt that method for the following reasons: 1) the average number of collaborators is small. 2); the funding body cannot apprehend real amounts of money allocated to collaborators.

After aggregating the numbers of proposals granted and amount of money awarded by university by discipline, radar charts for each university were drawn. The charts are common logarithmic-scale ones. In the figures shown below, 4 (outer circle) means No. 1, 3 (second-most outer circle) means 1/10th of the No. 1 university, and so on. We also produced tables of the top 30 universities in terms of the number of grants and amount of money.

Lastly, to gain an understanding of the characteristics of the indicators, they were compared with bibliometric indicators (their publication years are 2006-15). As mentioned above, only disciplines having the same names of categories (i.e., ASJC subject categories) of Scopus were considered.

Interim results

Overall situation of Grants-in-Aid for Scientific Research

Table 1 shows the overall situation of Grants-in-Aid for Scientific Research categorized by university. It shows the lists of top-30 universities in terms of number of adopted proposals and amounts of money rewarded. Generally, national universities predominate. Only three prefectural/municipal universities and four private universities are among top 30 in the number of proposals or amount of money. The general trend confirms that the two indicators are correlated to university size and that universities with medical schools earned more.

Table 1. Overall situation of Grants-in-Aid for Scientific Research.

Universities	*# of granted proposals*	*Universities*	*Amount of money rewarded*
Tokyo (n)	12,856	Tokyo (n)	180,000,000
Kyoto (n)	10,171	Kyoto (n)	112,000,000
Osaka (n)	9,050	Osaka (n)	87,868,382
Tohoku (n)	8,772	Tohoku (n)	83,010,460
Kyushu (n)	6,513	Nagoya (n)	54,367,770
Hokkaido (n)	5,956	Kyushu (n)	52,955,410
Nagoya (n)	5,683	Hokkaido (n)	50,017,010
Tsukuba (n)	3,954	Tokyo Tech (n)	39,163,400
Hiroshima (n)	3,655	Tsukuba (n)	28,623,610
Keio (p)	3,344	Keio (p)	24,419,540
Kobe (n)	3,302	Hiroshima (n)	21,913,130
Tokyo Tech (n)	3,102	Kobe (n)	21,636,950
Waseda (p)	2,822	Waseda (p)	18,654,530
Okayama (n)	2,798	Okayama (n)	16,980,150
Chiba (n)	2,781	Chiba (n)	16,611,000
Kanazawa (n)	2,516	Tokyo Med & Dent (n)	14,591,730
Niigata (n)	2,191	Kanazawa (n)	13,838,720
Tokyo Med & Dent (n)	2,123	Kumamoto (n)	12,882,400
Kumamoto (n)	2,090	Niigata (n)	10,955,950
Nagasaki (n)	1,999	Nagasaki (n)	10,320,040
Tokushima (n)	1,862	Tokushima (n)	10,039,390
Nihon (p)	1,750	Ehime (n)	9,225,230
Gunma (n)	1,568	NAIST	9,158,240
Shinshu (n)	1,565	Tokyo Met. (m)	8,671,910
Yamaguchi (n)	1,551	Shinshu (n)	8,520,180
Kagoshima (n)	1,449	Tokyo Agr. & Tech. (n)	8,316,990
Ehime (n)	1,426	Osaka City (n)	8,119,320
Ritumeikan (p)	1,422	Osaka Pref. (m)	7,602,530
Tokyo Met. (m)	1,328	Ritsumeikan (p)	7,587,830
Osaka City (m)	1,305	Yamaguchi (n)	7,528,770

n = national university, p = private university, m = municipal/prefectural university

Next, let us analyze the radar chart of the University of Tokyo, which is No. 1 in terms of numbers of proposals as well as amount of money (Figure 1). Even at the discipline level, indicators give high scores over the disciplines. Nevertheless, there are a few older disciplines (e.g., management and materials engineering) in which the University of Tokyo does not belong to the top group.

Distinctive discipline and university

The "brain sciences" discipline, which people often regard as a cutting-edge research field, is interesting. The "best" private university in this discipline is Tamagawa University, which keeps up with national research universities in the discipline (No. 6 in both number of grants and amount of money). The university started a graduate school in brain sciences in 2014 and has a relatively long history of research in the field; hence, it is worth noting it here for its

research and education. Its radar chart (Figure 2) shows that the university performed reasonably well in disciplines relevant to "brain sciences." That is, it performed well in human informatics, childhood science, linguistics, psychology, education, and neuroscience. Although its overall research capacity is not so large, the university seems to be an example in which "functional differentiation" has an adverse impact on local disciplinary agglomeration.

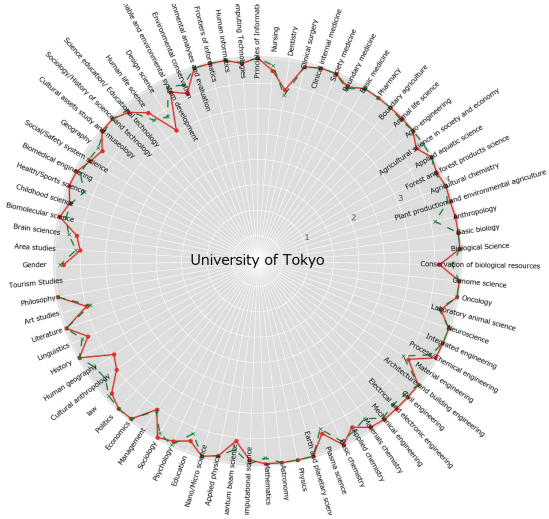


Figure 1. Number of granted proposals (dashed line) and amount of money rewarded to the University of Tokyo by research discipline (FY 2006-15).

Table 2. Funding-based indicators' correlations to bibliometric indicators.

R^2	*FWCI (Field Weighted Citation Impact)*	*number of papers*	*institutional h-index*
# of proposals granted in psychology	0.01	0.70	0.66
Amount of money rewarded in psychology	0.00	0.83	0.78
# of proposals granted in EPS	0.02	0.96	0.77
Amount of money rewarded in EPS	0.01	0.92	0.74

Note: The figure of 0.83, for example, indicates that R^2 between the amount of money rewarded in psychology and the number of papers in the field is 0.83.

Comparison with bibliometric data

To gain a better understanding of the characteristics of the indicators based on the Grants-in-Aid for Scientific Research data, we compared them with bibliometric indicators. In particular, we considered disciplines with the same names in Scopus's ASJC subject category system, "earth and planetary sciences", "mathematics", "nursing", "dentistry", "psychology", "neuroscience". Among them, we focused on disciplines that require more funding where researchers in many universities are doing research, i.e., "earth and planetary sciences (EPS)" and "psychology". Table 2 lists the correlations (R^2) between the funding-based indicators and bibliometric indicators.

Although we should not generalize these preliminary results, the funding-base indicators seem to have little correlation to the impact indicators. Meanwhile, they are strongly correlated to the number of papers as well as the institutional h-index; thus, they seem to be size-dependent.

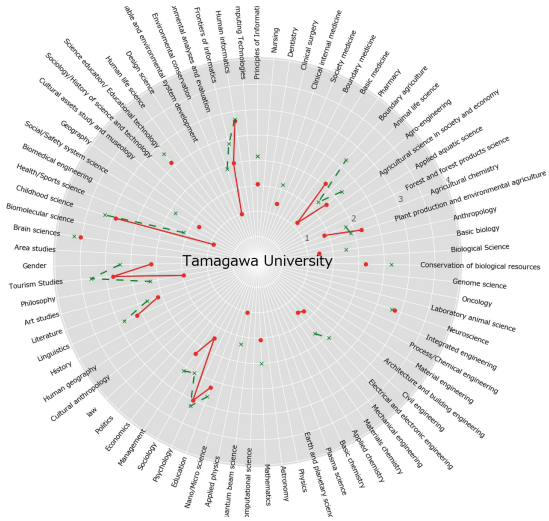


Figure 2. Number of granted proposals (dashed line) and amount of money rewarded to Tamagawa University by research discipline (FY 2006-15).

Discussion and future work

The above results confirm that funding-based indicators, which are aggregated at the detailed discipline level, are useful for measuring the research capacity of universities while at the same time respecting the diversity of university research activities. Preliminary analysis of correlations to bibliometric indicators, however, suggests that the significance of the indicators have to be considered carefully, because they are strongly correlated to the number of papers. So, it is necessary to identify the structure of both funding-based and bibliometric indicators by performing statistical analyses like PCA. For that to be possible, journal articles in the bibliometric database will have to be re-categorized.

Acknowledgments

This work was supported partly by JSPS (JSPS KAKENHI Grant Number 16H06580 & 17K01173). The bibliometric data used in this paper were processed from the Scopus/SciVal database with the cooperation of Elsevier. The author thanks Dr. Amane Koizumi at NINS for having a useful discussion.

References

Hicks, D., Wouters, P., Waltman, L., de Rijcke, S. & Rafols, I. (2015). Bibliometrics: The Leiden Manifesto for research metrics. Nature, 520, 429–431.

Koizumi, A. (2016). *Basic research on development of indicators to measure research capacity (by subjects / by university functions) and visualization of institutional research capacity*. Retrieved March 7, 2017 from: http://runetwork.jp/activity/detail.php?id=22. (in Japanese).

Koizumi, A. & Shirabe, M. (2017). How we measure research performance in Japanese universities?. Hitotsubashi Business Review, Jun, 58-72. (in Japanese)

Phillips, N. (2017). Striving for a research renaissance. Nature, 543, S7.

Shirabe, M. (2017). World University rankings and “research capacity”. IDE, 589, 45-51. (in Japanese)

Retracted Research Articles from the RetractionWatch Data Base Research in Progress Paper

Judit Bar-Ilan[1] and Gali Halevi[2]

[1]*Judit.Bar-Ilan@biu.ac.il*
Department of Information Science, Bar-Ilan University, Ramat Gan, 5290002, Israel

[2] *gali.halevi@mssm.edu*
Icahn School of Medicine at Mount Sinai, New York, NY, 10029

Abstract

This paper reports on the main characteristics of a selected sub-set of the newly launched RetractionWatch database. The database contains over 17,000 retracted articles and retraction notifications across various disciplines. In addition to bibliographic data such as title, journal name, authors and dates, the database also includes the retraction dates and reasons for retraction, the publisher and subject areas of the retracted articles. Our analysis includes the type of retractions, the time lag between publication to retraction as well as the main reasons for the retraction. In addition to these characteristics we also included some observations of the citations counts of pre- and post- retractions.

Introduction

The peer review process is a safe guard measure that attempts to ensure that the science reported in articles is vetted for accuracy, reproducibility and credibility. However, despite of it, there are articles that are retracted post publication once the editor/s or readers discover inaccuracies, errors and other issues that could not be amended prior to publication or could not be corrected in the published work. Retractions could occur due to various reasons including ethical reasons (e.g. plagiarism or authorship disputes), scientific errors or misconduct (e.g. image/data fabrication, intended/unintended mistakes in analysing results) and for administrative reasons (e.g. article published in the wrong issue) (see Bar-Ilan & Halevi, 2018). The topic of retracted articles has been studied through the years because of the gravity and negative impact it has on science especially when the reason for retraction is erroneous data and analytics, misconduct or unethical behaviour (e.g. Grieneisen & Zhang, 2012; Fanelli 2013; Lu, Jin, Uzzi, & Jones, 2013; van Leeuwen & Luwel, 2014; Almeida, de Albuquerque Rocha, Catelani, Fontes-Pereira, & Vasconcelos, 2015; Bar-Ilan & Halevi, 2017 & 2018). The retraction process itself is not necessarily straightforward and certainly not swift. There are many articles that are retracted years after they were published when errors, misconduct or deliberate manipulation of data are discovered. This mainly means that these articles receive citations throughout the duration of the publication cycle. However, studies have shown that even after retraction, some articles continue to be cited despite of retraction notices being posted by publishers (see Unger and Couzin 2006; Campanario 2000; Redman, Yarandi, & Merz, 2008; Bar-Ilan & Halevi, 2017 & 2018). These citations are categorically called "post retraction citations".

According to the Retraction Guidelines of the Committee on Publication Ethics (COPE, 2009), notices of retracted articles should include links to all electronic versions of the article, clearly identify the retracted article and have to be clearly identified as a retraction (COPE, p. 2). However, with the availability of pre-prints and post-print versions on the Web (e.g. pre or post-print when green open access is allowed) there are many cases where the retraction notices are not assigned to all such version and the article, therefore, continues to be cited as a genuine article (Bar-Ilan & Halevi, 2017). The RetractionWatch blog by Ivan Oransky and Adam

Marcus (Wikipedia Contributors, 2018), has been monitoring retracted articles for several years. Recently, they have been curating a list of retracted articles in order to ensure that retracted articles are easily identified. In addition to the standard metadata (title, journal, authors and their affiliation), they also provide detailed information on publication and retraction dates, DOIs/PubMed IDs (where available) of both the retracted item and the retraction notice, the reason for retraction and the discipline the article is categorized under.
Based on our previous studies on retractions, we were granted access to the RetractionWatch dataset in July 2018. At that time, the database contained more than 17,500 records with approximately 8,000 records labelled as "Research Article". Since the database also includes the publication and the retraction date, we were able to differentiate between *pre*- and *post*-retraction citations and offer some insights into the phenomena of *post*- retraction citations.

Data collection

The data we received from RetractionWatch, in July 2018, contained 17,679 records overall. Out of the entire dataset, 7,928 items were identified as "Research Article". In order to be able and track citations to these articles, we first retrieved articles that had a DOI or a PubMed ID and had both a publication date as well as a retraction date. In addition, we selected articles published in or after 2010 as we wanted to study more recent retracted articles. Once all these conditions were met, our final dataset contained 4,372 records. Our initial examination of the dataset showed that there were some errors with the dates assigned to either the publication, retraction date or both. For example, we found 230 cases of research articles where the retraction date was the same or earlier than the publication date. In addition, we also found four duplicates and 17 records with date of publication before 2010 which were removed from the analyzed dataset. Finally, we used Scopus to track citations to the pre- and post- retracted article. Since 686 articles were not indexed in Scopus they were removed, thus leaving 3,435 retracted research articles in our dataset. These articles were all published on or after 2010, had DOIs or PubMed IDs, contained publication and retraction dates and were indexed by Scopus. Since the main aim of this study was to characterize the dataset and assess the amount of citations received by an article pre- and post- its retraction, we used Scopus to retrieve the yearly citation counts for the article before and after it was retracted. Citation data from Scopus were collected in the last week of December 2018 and the first week of January 2019. At the end of the data collection and cleansing processes, the dataset contained 3,435 retracted articles which were further analyzed.

Results and Discussion

The time span between publication and retraction is a very important topic in the study of retracted articles and their citations. The longer it takes for an article to be retracted and for a notice to be published, the more citations it may potentially receive. We first examined the dataset in terms of periods of peak publications and retractions. As can be seen in figure 1, in 2014 there seemed to be a peak in the number of articles published in or after 2010 that were later retracted. Accordingly, in 2016 there seems to be a peak in the number of retractions notices. This leads us to believe that there is a 2-year time span between publication and retractions. That said, we did notice that the average time between publication and retraction keeps decreasing (see Figure 2).

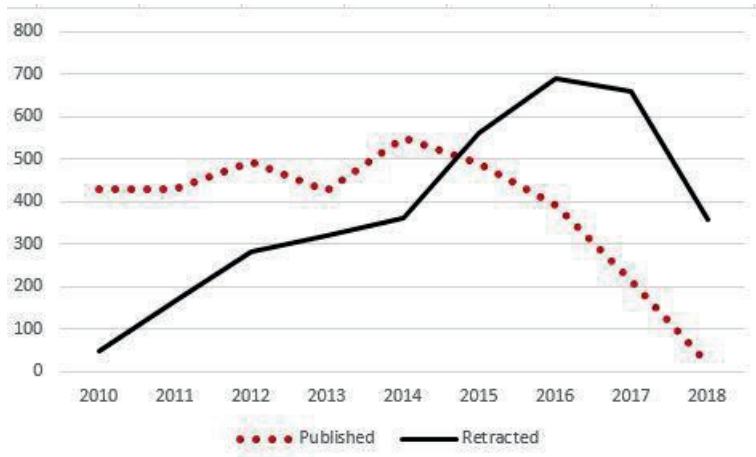


Figure 1. Yearly distribution of articles retracted alter, and the number of retractions per year

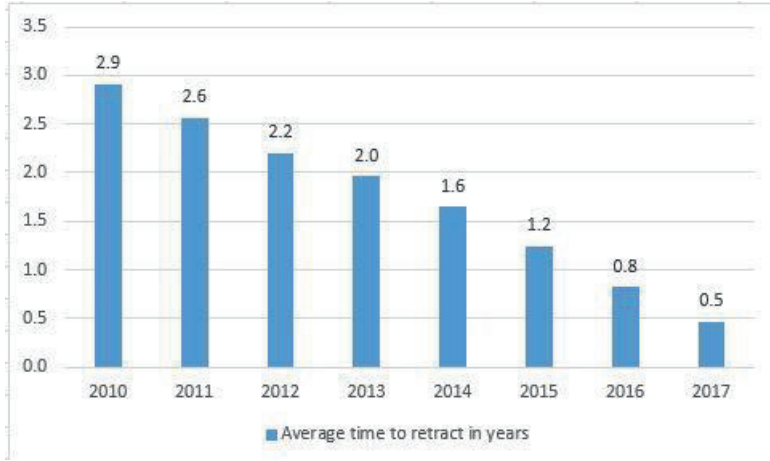


Figure 2. Average time to retract in year per year of publication

RetractionWatch includes reasons for retraction and subject categories for each of the articles in the database. In many cases there is more than one reason stated or subject category chosen. In order to gain some insight into the most recurring reasons for retractions we aggregated the prevalent ones (see table 1). As can be seen in table 1, the two most prevalent reasons for retractions are *Scientific Misconduct* and *Plagiarism/Duplication* which represents 34.8% and 24.8% of the reasons listed respectively. Interestingly, there were also four cases where the reason for retraction was that article cites previously retracted work. We analysed subject categories the database assigns to each article in order to identify the disciplines where retractions occurred the most. In line with previous studies, retractions predominantly occur in Life and Biological Sciences and Health Sciences. (see table 2). The RetractionWatch database also contains the journal and publisher information for each of the retracted article. Our dataset contained 225 publishers overall. Figure 3 displays the publishers with 50 or more retracted articles in our dataset. Our data shows that 12 publishers cover 72.8% of the retracted articles.

It is not surprising that Elsevier and Springer are among the top publishers with the largest number of retracted articles. As shown in table 2, the majority of retracted articles are life and medical sciences related and since Elsevier and Springer are leading publishers in these disciplines, they also have the largest amount of articles' retractions. Figure 4 displays the names of fifteen journals with the largest number of retracted articles.

Post retraction citations were defined as any citation received a year or more after the retraction year. As stated above, yearly citation data were retrieved from Scopus by searching the articles DOIs on Scopus and using the "View citation overview" option on the platform. Table 3 shows that on average retracted articles receive less post retraction citations than pre-retraction citations, which is encouraging. This means that articles are cited in much lower rate once retracted. However, for one third of the articles, the number of post retraction citations is greater than the number of pre-retraction citation. Although the differences are relatively small, it is still a phenomenon that needs to be noted. The largest difference we noticed, with 96 post retraction citations out of 118 total citations was received by the paper *"A Glucagon-like Peptide-1 (GLP-1) Analogue, Liraglutide, Upregulates Nitric Oxide Production and Exerts Anti-inflammatory Action in Endothelial Cells"* by Hattori et al., in the journal Diabetologia, published in 2010 and retracted in 2011. The article that received the largest number of post retraction citations (125 citations) is the *"Long term toxicity of a Roundup herbicide and a Roundup-tolerant genetically modified maize"* by Séralini et al., published in 2012 in Food and Chemical Toxicology, and retracted in 2013. It should be noted that the article was republished "as is" in Environmental Science Europe in 2014 (Fagan, Traavik, & Bøhn, 2015). The republished article received 84 citations so far.

Table 1. Reasons given for retraction

Reasons for retraction	Frequency	Percentage
scientific misconduct	2864	34.8%
plagiarism/duplication	2044	24.8%
investigation	1103	13.4%
authorship issues	677	8.2%
publication ethics	435	5.3%
fake peer review	403	4.9%
other	153	1.9%
error by journal/publisher	140	1.7%
no reason given	132	1.6%
withdrawal	93	1.1%
referencing	90	1.1%
legal	38	0.5%
author unresponsive	32	0.4%
miscommunication	25	0.3%
cites prior retracted work	4	0.0%

Table 2. Retracted articles by subject

Subject category	# of articles	Percentage
Life and Biological Sciences	1,839	53.5%
Health Sciences	1,330	38.7%
Physical Sciences: Mathematics, Physics, Chemistry, Chemical Engineering	1,018	29.6%
Business/Technology	364	10.6%

Environmental Sciences	265	7.7%
Social Sciences	192	5.6%
Humanities	26	0.8%

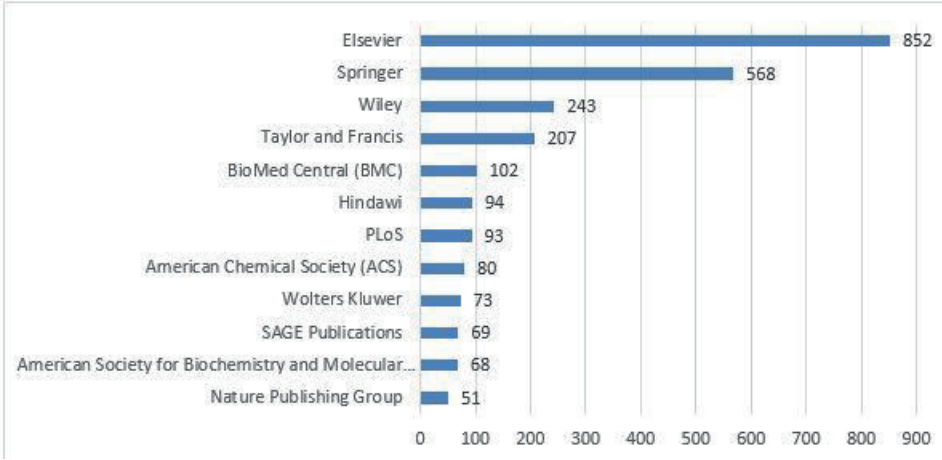


Figure 3. Publishers with the largest number of retracted articles

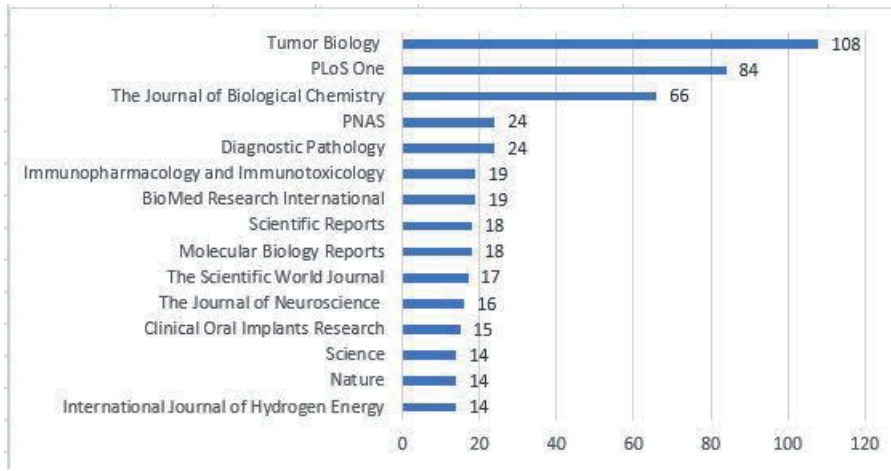


Figure 4. Journals with the largest number of retracted articles

Table 3. The average number of pre and post retraction citations by year of publication

Pub year	Average # of citations	Average # of citations before retraction	Average # of citations after retraction
2010	21.7	14.2	7.5
2011	21.5	14.3	7.2
2012	13.4	8.2	5.2
2013	18.8	14.5	4.3
2014	9.2	5.8	3.4
2015	5.9	3.9	2.0
2016	5.4	3.5	1.8
2017	3.6	2.9	0.7
Total	12.7	8.6	4.2

Conclusion

This study is based on a subset of articles included in the RetractionWatch database. For the purpose of this study we chose a subset of articles that were published in or after 2010 and had a DOI or a PubMed ID and publication and retraction dates so we could analyse the citations an article received pre- and post- retractions. Our analysis showed that the time between publication and retraction decreases and now averages 2 years or less. Although we cannot be certain as to the reason for the decrease, it may be the rapid manner by which articles are electronically published nowadays thus reaching and scrutinized by their audience much faster than before. Another related reason could be the availability of software tools that detect issues such as plagiarism much faster than before. We also believe that the increasing availability of open data makes reproducing results much easier and contributes to articles being scrutinized and criticized for data inaccuracies and misconducted analysis. Much in line with other studies, we also found that life and medical sciences have much more retractions than any other subjects. Since these topics account for over 25% of the overall published scientific output ("S&E Indicators 2018 | NSF - National Science Foundation," n.d.), it is not surprising that the amount of retractions within these disciplines will be the highest. On the same token, Elsevier and Springer are the publishers with the most retractions assigned to them. This correlates to the above life and medical sciences retractions rates as these two publishers hold the largest number of journals in these areas. Our data show that the average citations of articles is significantly lower post- retraction, but it is still concerning that they persistently cited even after the retraction notice has been issued. There could be several reasons that could potentially explain this. First, there might be versions of the original article still in circulation; deposited to repositories or hosted on accessible websites, these articles could be retrieved and cited since the official publisher retraction notice will not display in these outlets. Pre and post prints are often uploaded, shared and deposited and while the official journal might issue a retraction notice, these websites would still host the original article. Secondly, there are times when retracted articles are 'negatively' cited (Bar-Ilan & Halevi, 2018). Although they are negatively cited within the context of the article, the citation is still recorded and assigned to the original article. Finally, some publishers make retracted articles free for readers. On the one hand, this could be a positive thing, whereas readers can be aware of retracted research. On the other hand, this free information can be used without limit which can create not only citations but also increased readership, downloads and shares of an erroneous science. COPE Guidelines (2009, p. 2) do not allow to remove retracted articles, however hopefully the RetractionWatch database will allow authors to refrain from citing retracted articles and enable editors/publishers to check that retracted articles are not cited.

Acknowledgments

We thank RetractionWatch for providing the data on retracted articles.

References

Almeida, R. M. V., de Albuquerque Rocha, K., Catelani, F., Fontes-Pereira, A. J., & Vasconcelos, S. M. (2015). Plagiarism allegations account for most retractions in major Latin American/Caribbean databases. *Science and Engineering Ethics, 22*, 1447–1456.

Bar-Ilan, J., & Halevi, G. (2017). Post retraction citations in context: a case study. *Scientometrics, 113*(1), 547-565.

Bar-Ilan, J., & Halevi, G. (2018). Temporal characteristics of retracted articles. *Scientometrics*, *116*(3), 1771-1783.

COPE (2009). Retraction Guidelines. doi: 10.24318/cope.2019.1.4

Campanario, J. M. (2000). Fraud: retracted articles are still being cited. *Nature, 408*(6810), 288.

Grieneisen, M. L., & Zhang, M. (2012). A comprehensive survey of retracted articles from the scholarly literature. *PLoS ONE, 7*(10), e44118. doi: 10.1371/journal.pone.0044118

Bar-Ilan, J., & Halevi, G. (2017). Post retraction citations in context: a case study. *Scientometrics*, *113*(1), 547-565.

Fagan, J., Traavik, T., & Bøhn, T. (2015). The Seralini affair: Degeneration of science to re-science? *Environmental Sciences Europe*, *27*(1), 19.

Fanelli, D. (2013). Why growing retractions are (mostly) a good sign? *PLoS Medicine, 10*(12), e1001563. doi:10.1371/journal.pmed.1001563.

Li, G., Kamel, M., Jin, Y., Xu, M. K., Mbuagbaw, L., Samaan, Z., ... & Thabane, L. (2018). Exploring the characteristics, global distribution and reasons for retraction of published articles involving human research participants: a literature survey. *Journal of multidisciplinary healthcare*, *11*, 39.

Lu, S. F., Jin, G. Z., Uzzi, B., & Jones, B. (2013). The retraction penalty: Evidence from the Web of Science. *Scientific Reports, 3*, 3146.

Redman, B. K., Yarandi, H. N., & Merz, J. F. (2008). Empirical developments in retraction. *Journal of*
Medical Ethics, 34(11), 807–809.

S&E Indicators 2018 | NSF - National Science Foundation. (n.d.). Retrieved from https://www.nsf.gov/statistics/2018/nsb20181/

Unger, K., & Couzin, J. (2006). Even retracted papers endure. *Science, 312*(5770), 40–41.

van Leeuwen, T., & Luwel, M. (2014). An in-depth analysis of papers retracted in the Web of Science. STI 2014 Leiden, 337.

Wikipedia contributors. (2018). Retraction Watch. In *Wikipedia, The Free Encyclopedia*. Retrieved from https://en.wikipedia.org/w/index.php?title=Retraction_Watch&oldid=862414126

Open Science Behavior of AI Industry: Collaboration Patterns and Topics from the Perspective of Cross-Institutional Authors

Xiaoling Sun[1], Kun Ding[1] , Yuan Lin[1] and Hongfei Lin[1]

xlsun@dlut.edu.cn, dingk@dlut.edu.cn, zhlin@dlut.edu.cn, hflin@dlut.edu.cn
[1] Dalian University of Technology, Dalian (China)

Abstract

In this study, the open science behaviour of industry is analyzed in terms of publications co-authored by at least one author from industry. Firstly, authors are classified into five types according to affiliations, and collaboration network is built to investigate the role of different types of authors. Then, topic evolution of papers related to industry is studied using Hierarchical Dirichelet Process and topic mapping algorithm. Artificial intelligence research area is taken as an example. Results show cross-university-industry type authors play an important bridge role in the collaboration network. And research topics that industries are concerned about are revealed and analyzed. This research could provide reference for formulating science and technology policy and promoting scientific research innovation in both universities and industries.

Introduction

In the global context of open innovation and science-based innovation, industries are more and more frequently taking the action of open science, voluntarily disclosing innovative knowledge rather than confidentiality (Simeth and Lhuillery, 2015). Large industries with independent laboratories (such as AT&T, IBM, Google, Tesla) and many small and medium-sized technology-based industries are actively taking the action of open science (Åstebro et al., 2012). Especially in recent years, the development of biopharmaceutical industry, nanotechnology industry, new energy, new materials and other emerging industries is increasingly dependent on scientific innovation. The scientific capabilities of industries include not only R&D capabilities (Azagra et al., 2019), but also scientific disclosure capabilities (Jong and Slavova, 2014). Hicks (1995) found that industries regularly publish papers in scientific journals, such as Philips, Siemens, Hitachi, Sanders and so on. The number of papers published by large industries is even equal to that of a medium-sized research university.

University-industry collaboration (UIC) is a popular way that industries publish papers, in which industries and universities could take full advantage of their own strengths. University-industry relations have attracted growing interests of the research community. The research community has tried to find why do universities and industry collaborate and how to make this relation more efficient from different perspectives, such as technology transfer (Etzkowitz et al., 2000; Gulbrandsen and Slipersaeter, 2007; Chai and Shih, 2016), government roles (Leydesdorff and Etzkowitz, 1996; Ponomariov, 2013), and individual motives (Ankrah et al., 2013; Callaert et al., 2015). Tijssen (2012) studied publications co-authored by university-industry collaboration, and generated university-industry research connections (UIRC) data. Wong and Singh (2013) found a positive effect of university-industry collaboration on the commercialization of university technology. Zhou et al. (2016) analyzed university-industry collaborations in China and the USA in terms of co-authored publications indexed in the Web of Science (WoS), and results showed a wide gap between China and the USA.

Though there are many research related to open science behaviour of industry and UIC, there is little research that focuses on the cross-institutional authors from collaboration patterns and topics. Hottenrott and Lawson (2017) studied the phenomenon of multiple institutional affiliations using scientific publications and found that multiple affiliations had at least doubled over the past few years. We also found that some authors engaged in both university and industry. *What is the role they play in the science and technology relationship*? *What research topics they are interested in?* In general, this part of knowledge belongs to the Pasteur's quadrant,

focusing on “use-inspired basic research” with "science-technology" dual value (Murray,2002). Driven by these questions, we performed the studies in the field of artificial intelligence (AI). In recently years, AI-related technologies penetrate almost all fields (Esteva et al., 2017; Arel et al., 2010), and new scenarios have been activated in many industries, bringing new values and attracting worldwide investment and attention.

The rest of this paper is organized as follows: Section *Data and Methods* introduces the dataset and the proposed method including how to identify different types of authors and how to characterize collaboration patterns and topics; Section *Results* presents the results of the proposed method; then we conclude in Section *Conclusions*.

Data and Methods

Dataset

Publication data related to artificial intelligence is derived from Web of Science (WoS) based on the search strategy: TI = ("artificial intelligence" OR "deep learning" OR "neural network*" OR "machine learning" OR "machine Intelligence" OR "face detection" OR "face recognition" OR "speech recognition" OR "image recognition" OR "expert system*" OR "artificial life"). The data covers the period 1945-2017 and there are totally 163,713 publications. As in this study we concern about the open science behaviour of industry, the publications that have at least one author from industry are considered.

Identifying different types of authors

Author affiliations of publications indexed in WoS make it possible to quantitatively analyze institutional engagement in university-industry collaboration. According to the author affiliations, we classify authors into five types: *U(one university affiliation)*, *UU(multiple university affiliations)*, *UI(multiple university and industry affiliations)*, *II(multiple industry affiliations)*, *I(one industry affiliation)*. *UI*, *UU*, and *II* authors are cross-institutional authors who engage in more than one institution, while *UI* authors engage in both university and industry (e.g. Fig.1).

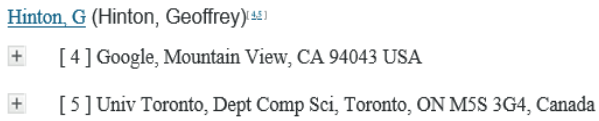


Figure 1. An exemplary author engages in both university and industry.

In order to accurately identify university and industry, several rules are used to identify institutions. If the affiliation name contains ("* corp" or "* inc" or "* co ltd" or "* llc" or "* co"), it is an industry, and if the name contains ("*univ*" or "* coll" or "* lab" or "* res inst" or "* res ctr" or "*r&d"), it is an university or research institution. The rules cannot identify all affiliation names, therefore, three graduated students participated in the annotation task to label the unidentified affiliations. After the labelling task, authors are classified according to the affiliations.

Characterizing collaboration patterns

University-Industry collaboration network is built based on author co-authorships (Fig. 2), in which authors are nodes and the co-author relationships among them are edges. The weight of the edge is the number of times two authors collaborate. Based on the collaboration network, Pagerank algorithm (Sergey and Lawrence, 1998) is used to identify the key authors in

university-industry collaborations. Pagerank is a way of determining influence within a network, and was initially used to sort search results.

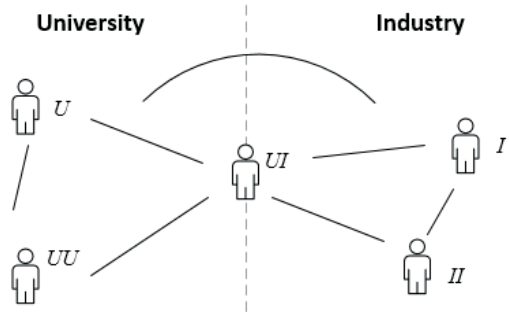


Figure 2. University-Industry collaboration network from the perspective of authors.

Unveiling topic evolution

Firstly, based on the Hierarchical Dirichlet Process (HDP) model, the publication keywords are clustered to get the number of topics and the distribution of topics automatically. HDP model is a non-parametric Bayesian topic model (Teh, 2006). Compared with the classical LDA model, it has the advantage of automatically determining the number of topics according to the data set and is widely used in many fields, especially in text mining.

Based on HDP algorithm, the research topics in different periods are revealed. Next, we need to map the topics in different periods and explore their changing trends. The most direct way to distinguish the differentiation, fusion, generation or disappearance of topics is based on the overlap of topics. Topics are composed of keywords, and the higher the degree of keyword overlap, the greater the similarity between two topics. Here Jaccard similarity is used to compute the similarity as shown in Eq. 1.

$$C_A(t) = \frac{|\mathrm{A}(t_0) \cap \mathrm{A}(t_0 + t)|}{|\mathrm{A}(t_0) \cup \mathrm{A}(t_0 + t)|} \tag{1}$$

Where t is the time interval, and A(t_0) represents the topic at time t_0. The similarity between two topics is calculated in two adjacent time periods. When the similarity is greater than a threshold, the two topics are considered to be the same topic in the adjacent time periods. If a topic is similar to multiple topics in the next period, the topic becomes divided; if multiple topics are similar to the same topic in the next period, the topics are merging; if a topic has no similar topic in the previous period, it can be considered as a newly generated topic; if a topic has no similar topic in the next period, it disappears.

Results

The number of AI publications and authors in every year from 2008 to 2017 could be seen from Fig. 3(left). In general, the data could be divided into three stages: 2008-2011, 2012-2014 and 2015-2017. The top 25 disciplines the publications belong to are shown in Fig. 3(right), which cover a wide range of research areas.

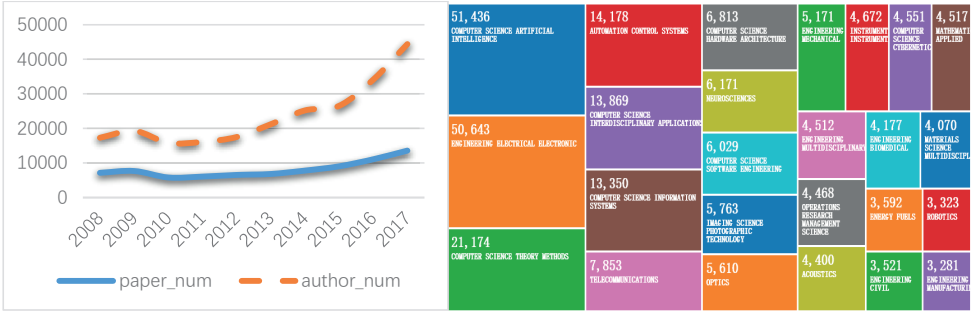


Figure 3. Number of publications and authors in every year (left); Top 25 Web of Science categories (right).

Cross-institutional authors

According to the identification of the types of authors, the ratio of different types of authors is shown in Fig. 4. Cross-institutional authors have a relatively high proportion of all authors, especially *UI* authors, which are increasing during three stages. The authors from university have the highest proportion, as the most popular university-industry collaboration way is authors from industry collaborate with authors from university, maximizing their own strengths.

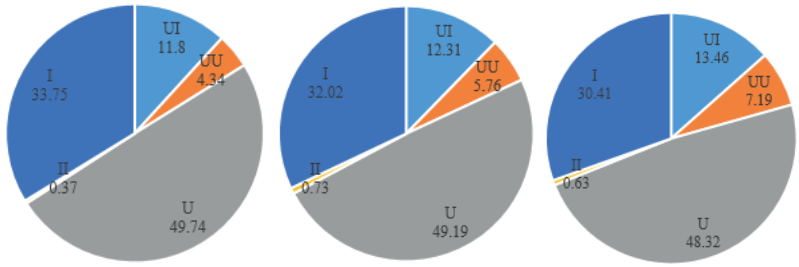


Figure 4. The ratio of different types of authors in three stages.

Table 1 lists top 30 industries and universities that engaged in university-industry collaboration in three stages. IBM, Microsoft, NTT, Google, et al. are the leading industries that are active in open science behaviour in the AI research field. Univ Tokyo, Chinese Acad Sci, Univ Calif Los Angeles, et al. are the leading universities that are active in coauthoring with industry. The rank is changing in different stages.

Collaboration patterns

The collaboration networks are built in three stages and pagerank algorithm is used to identify important nodes. In university-industry collaborations, *UI* type authors account for 11.8-13.5%. It is not very surprised that they play an important bridge role in the UIC, which could be reflected by the results in table 2. *UI* type authors have a large proportion of the top 20 authors, which are from 45% to 60%.

The largest component of collaboration network of the first stage (2008-2011) is visualized in Fig. 5. There are four *UI* authors: Sakai Makoto [DENSO Corp, Nagoya Univ], Kolossa Dorothea [Tech Univ Berlin, NTT Corp, Univ Calif Berkeley], Keutzer Kurt [Synopsys Inc, Univ Calif Berkeley, Universal Parallel Comp Res Ctr], and You Kisun [Seoul Natl Univ, Intel Corp]. They occupy important structural holes in the collaboration network.

Table 1. Top 30 industries and universities that engaged in university-industry collaboration.

Rank	2008-2011		2012-2014		2015-2017	
	Industry	**University**	**Industry**	**University**	**Industry**	**University**
1	IBM Corp	Univ Tokyo	DESY	CERN	Google Inc	Chinese Acad Sci
2	CNR	Chinese Acad Sci	Microsoft Corp	Ist Nazl Fis Nucl	NTT Corp	Tsinghua Univ
3	Microsoft Corp	Nagoya Inst Technol	IBM Corp	CNRS	IBM Corp	Nanyang Technol Univ
4	NTT Corp	Univ Calif San Diego	NTT Corp	Rhein Westfal TH Aachen	CNR	Radboud Univ Nijmegen
5	Japan Sci & Technol Agcy	Seoul Natl Univ	Google Inc	Chinese Acad Sci	Mayo Clin	Natl Univ Singapore
6	GEMTEX	ENSAIT	CNR	CEA Saclay	Intel Corp	Stanford Univ
7	Henry Ford Hosp	Univ Lille	CEA	Univ Bonn	Samsung Elect	Huazhong Univ Sci & Technol
8	ETRI	Bogazici Univ	UPMC	Joint Inst Nucl Res	Google	Harvard Med Sch
9	Mayo Clin	Georgia Inst Technol	OeAW	Fermilab Natl Accelerator Lab	Microsoft Corp	Univ Michigan
10	NASA	Univ Calif Berkeley	Google	Univ London Imperial Coll Sci Technol & Med	NASA	Carnegie Mellon Univ
11	Toshiba Res Europe Ltd	Univ Grenoble 1	Ecole Mines Ales	Univ Michigan	Massachusetts Gen Hosp	Korea Adv Inst Sci & Technol
12	NHK Japan Broadcasting Corp	MINES Paris Tech	INFN	Univ Valencia	Google DeepMind	Univ Toronto
13	Adv Technol Applicat Ctr	Univ Lausanne	Microsoft	MIT	IBM Res Almaden	Imperial Coll London
14	Intel Corp	Univ Michigan	Boys Town Natl Res Hosp	Univ Calif San Diego	Microsoft	Univ Cambridge
15	Numerate Inc	Toyama Univ	Kaiser Permanente	Univ Calif Berkeley	ETRI	MIT
16	Samsung Elect Co Ltd	Soochow Univ	LIMSI CNRS	Univ Freiburg	Brigham & Womens Hosp	Univ Tokyo
17	Hermanus Magnet Observ	INSERM	Natl Def Med Ctr	Univ Amsterdam	NVIDIA	Univ Calif Los Angeles
18	ERATO	Harvard Univ	Fdn Bruno Kessler	Univ Karlsruhe	US EPA	Seoul Natl Univ
19	Ford Motor Co	Griffith Univ	NASA	NIKHEF H	Mind Res Network	Minist Educ
20	Shandong Inst Light Ind	JST	NEC Corp Ltd	Swiss Fed Inst Technol	Toshiba Co Ltd	Univ Edinburgh
21	CEP TEP Ctr Energet & Proc	Osaka Univ	Massachusetts Gen Hosp	Univ Wisconsin	Educ Minstry China	Minist China
22	Siemens AG	Univ W Bohemia	Taipei Vet Gen Hosp	Univ Florida	Fdn Bruno Kessler	Northeastern Univ
23	IBM Japan Ltd	Tokyo Inst Technol	Spansion Inc	Univ Bologna	Microsoft AI & Res	Johns Hopkins Univ
24	Inst Nacl Invest Nucl	Univ Milan	Fujitsu Res & Dev Ctr Co Ltd	Rutherford Appleton Lab	CEA	Hubei Normal Univ
25	Pfizer Inc	Northeastern Univ	Brigham & Womens Hosp	Tel Aviv Univ	Facebook AI Res	Tel Aviv Univ
26	NEC Labs Amer Inc	Univ Bologna	Singapore Gen Hosp	Univ Iowa	MC10 Inc	Univ Sydney
27	Mind Res Network	Nanyang Technol Univ	Natl ICT Australia	Univ Tokyo	Cambricon Technol Co Ltd	Univ Washington
28	HUGE	Univ London Imperial Coll Sci Technol & Med	Res Ctr Food Fermentat Engn & Technol Hubei	Univ Manchester	Childrens Hosp Philadelphia	Harvard Univ
29	CEA	Florida Int Univ	PPNS	Harvard Univ	Adobe Res	Univ Illinois
30	SpeechTech S R O	McMaster Univ	Toshiba Co Ltd	Heidelberg Univ	Panasonic Corp	Georgia Inst Technol

Table 2. Top 20 authors in University-industry collaboration networks ranked by pagerank.

Id	2008-2011			2012-2014			2015-2017		
	Name	Type	Pagerank	Name	Type	Pagerank	Name	Type	Pagerank
1	Aihara, Kazuyuki	***UI***	0.0012	Ney, Hermann	***UI***	0.0009	Zeng, Zhigang	***UI***	0.0006
2	Okada, Masato	***UI***	0.0007	Senior, Andrew	*I*	0.0008	Li, Haizhou	***UI***	0.0005
3	Fogel, Gary B.	*I*	0.0007	Heigold, Georg	*II*	0.0008	Zurada, Jacek M.	***UI***	0.0005
4	Grossi, Enzo	***UI***	0.0007	Ramabhadran, Bhuvana	*I*	0.0007	Ekins, Sean	***UI***	0.0004
5	Villa, Alessandro E. P.	***UI***	0.0006	Lam, Kin-Man	***UI***	0.0007	van Ginneken, Bram	***UI***	0.0004
6	Mohammadi, Amir H.	***UI***	0.0006	Li, Jinyu	*I*	0.0007	Wang, Dong	***UI***	0.0004
7	Raj, Bhiksha	***UI***	0.0006	Mendez-Vazquez, Heydi	***UI***	0.0007	Chandra, Rohitash	***UI***	0.0003
8	Zimmermann, Hans-Georg	*I*	0.0006	Gao, Yongsheng	***UI***	0.0006	Suri, Jasjit S.	***UI***	0.0003
9	Gao, Yongsheng	***UI***	0.0006	Chien, Jen-Tzung	***UI***	0.0006	Yu, Dong	*UU*	0.0003
10	Cordon, Oscar	*II*	0.0006	Sainath, Tara N.	*I*	0.0006	Xiao, Xiong	*U*	0.0003
11	Ahmad, Shandar	***UI***	0.0006	Gong, Yifan	*I*	0.0006	Senior, Andrew	***UI***	0.0003
12	Li, Yu-Chuan	***UI***	0.0006	Chen, Yan	***UI***	0.0006	Delcroix, Marc	*I*	0.0003
13	Kim, Byungwhan	*U*	0.0005	Singh, Rita	***UI***	0.0006	Lin, Liang	***UI***	0.0003
14	Franco, Annalisa	***UI***	0.0005	Johannet, Anne	*I*	0.0005	Bahadori, Alireza	***UI***	0.0003
15	Mendez-Vazquez, Heydi	*I*	0.0005	Schlueter, Ralf	*U*	0.0005	Beaufays, Francoise	*I*	0.0003
16	Cessac, Bruno	***UI***	0.0005	Liu, Xun	*UU*	0.0005	Yumer, Ersin	*I*	0.0003
17	Park, H. I.	*I*	0.0005	Zweig, Geoffrey	*I*	0.0005	Seltzer, Michael L.	*I*	0.0003
18	Lu, Hui	***UI***	0.0005	Liu, Jianzhuang	***UI***	0.0005	Wang, Yu	*UU*	0.0002
19	Lee, Chin-Hui	*UU*	0.0005	Niamsup, P.	***UI***	0.0005	Munteanu, Cristian R.	*UU*	0.0002
20	Khodayari-Rostamabad, Ahmad	*U*	0.0005	McDermott, Erik	*I*	0.0005	Li, Jinyu	*I*	0.0002
UI %	**60%**			**45%**			**55%**		

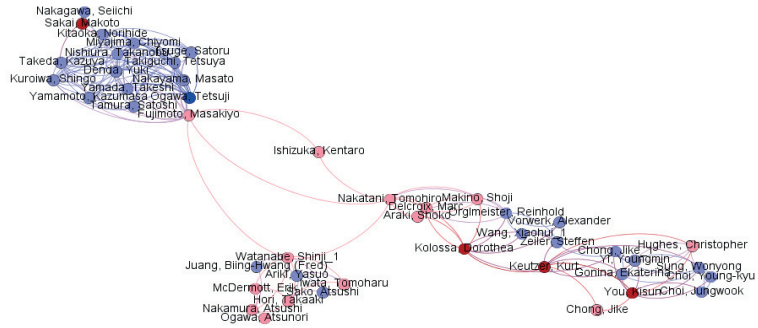


Figure 5. The largest component of collaboration network of the first stage (2008-2011). Red nodes (*UI*), pink nodes (*I*), blue nodes (*UU*), and light blue nodes (*U*).

Collaboration topics

The keywords in university-industry collaboration papers and the papers published by *UI* type authors are listed in Table 3, from which we could roughly see what issues industries and cross-university-industry authors pay close attention to. There was no doubt that the keywords that are used for data retrieval are the most popular ones. But still, some keywords such as "machine learning", "deep learning" and "neural network" have drawn much more attention recently. "classification", "prediction", "feature extraction" are the keywords with constant concerns in university-industry collaboration. Besides, there are also some issues of particular concerns to *UI* type authors, for example, "schizophrenia", "outlier detection".

Table 3. Top 20 keywords in UIC papers and *UI-author* papers.

2008-2011		2012-2014		2015-2017	
UIC papers	UI-author papers	UIC papers	UI-author papers	UIC papers	UI-author papers
neural network	neural network	neural network	artificial neural network	***machine learning***	***machine learning***
artificial neural network	artificial neural network	artificial neural network	neural network	***deep learning***	***deep learning***
speech recognition	speech recognition	machine learning	machine learning	***neural network***	***neural network***
machine learning	face recognition	speech recognition	speech recognition	artificial neural network	artificial neural network
face recognition	machine learning	face recognition	face recognition	convolutional neural network	convolutional neural network
artificial intelligence	**fmri**	deep neural network	**classification**	deep neural network	deep neural network
expert system	**classification**	**genetic algorithm**	**prediction**	speech recognition	speech recognition
classification	**feature extraction**	**prediction**	recurrent neural networks	recurrent neural network	recurrent neural networks
feature extraction	ann	**classification**	**genetic algorithm**	face recognition	**classification**
bp neural network	artificial intelligence	expert system	**fmri**	**classification**	automatic speech recognition
prediction	**schizophrenia**	bp neural network	**acoustic modeling**	automatic speech recognition	face recognition
ann	**nonwovens**	deep learning	automatic speech recognition	artificial intelligence	artificial intelligence
fuzzy neural network	**feature selection**	artificial intelligence	**pattern recognition**	**prediction**	**prediction**
genetic algorithm	expert system	face detection	artificial intelligence	cnn	**genetic algorithm**
automatic speech recognition	face detection	**modeling**	bp neural network	**support vector machine**	**data mining**
fmri	lvq neural network	automatic speech recognition	deep learning	**genetic algorithm**	**support vector machine**
image processing	**ionosphere**	robust speech recognition	**feature selection**	**data mining**	**long short-term memory**
optimization	**associative memory**	recurrent neural networks	**forecasting**	**feature extraction**	recurrent neural network
data mining	**outlier detection**	**acoustic modeling**	image processing	**feature selection**	**feature selection**
face detection	**support vector machines**	**optimization**	**optimization**	**lstm**	**schizophrenia**

In order to reveal the latent semantic information of document content, topic model is an effective method. For each stage's collection of documents, keywords are used as the subject content information of each document, and HDP model is used to analyze the topics of the document. The HDP model used in this paper is an open source algorithm with default parameters. Table 4-6 lists the top 10 topic distributions of AI in three stages. Each line is a topic, and the topic is represented by the top 10 keywords with the highest probability belonging to the topic.

From the tables, we can see the topics in each stage and the changes of topics over time. “neural network” related methods have always been the most popular topics in all stages. In recent years, “speech perception”, “complex network”, “gabor wavelets”, “autoencoders” et al. are also drawing increasing attention from industries in AI field.

Table 4. Topic Distribution of AI UIC research in 2008-2011.

Id	Topics
1	neural network, speech recognition, artificial neural network, machine learning, face recognition, artificial intelligence, feature extraction, bp neural network, prediction, genetic algorithm
2	artificial neural network (ann), sensitivity, fuzzy neural network, tokamaks, mechanical properties, cross validation, titanium alloy, syndrome distribution shaping, watershed, expert systems
3	expert systems, surface roughness, wavelet neural network, structural alerts, postictal state, robust bayesian neural network, reach, grain losses, color memory, error cause estimation
4	artificial neural networks (ann), self-organizing maps, microphone array, feature combination, elman neural network (enn), machining parameters, boolean series prediction questions (bspq), spiking neural networks, regioselectivity, expert systems
5	bottle-neck neural network, truecasing, decision support, high-throughput experimentation, fmri, connectivity, heart rate variability, fuzzy clustering, sentence boundary detection, auditory/visual evoked magnetic fields
6	fuzzy logic, groundwater abstraction impacts, modified probabilistic neural network, river-aquifer interactions, forest landscape composition and structure, case based reasoning, formal definitions,habitat modelling, adjustment tests, phase performance, synaptic depression
7	model adaptation, expert system, lvq neural network, pramipexole, galaxies evolution, prostatic neoplasm, lmi, genetically optimized hybrid fuzzy neural networks (ghfnn), cognitive engine (ce), mfcc
8	preferred firing sequence, rbf neural networks, apoptosis, synaptic pruning, adaptive control, artificial neural network, construction vehicle, neutron log, hematologic tests, autopilot
9	visual quality recognition, fingerprint, bayesian neural network, orthogonal activation function, outlier detection, regularization parameter, nonwovens, phase compensation, pattern matching, robust bayesian neural network
10	neural network (nn), very short term load prediction (vstlp), load dynamics, lawsuit, artificial neural network, modeling, isometry invariance, reverse vaccinology, hydrogen, wavelet transformation

Table 5. Topic Distribution of AI UIC research in 2012-2014.

Id	Topics
1	neural network, artificial neural network, machine learning, speech recognition, face recognition, genetic algorithm, classification, prediction, bp neural network, deep neural network
2	learning, uncertain dynamics, adaptive neural network (nn) control, adaptation, artificial neural networks (ann), persistent excitation (pe) condition, hearing aids, polycarbonate, reverberant speech recognition, wfst
3	neural network, diesel engine, exhaust gas recirculation, cluster analysis, mass air flow, indirect adaptive control, control, scale-invariant feature transform(sift), imbalanced data, reuse
4	robustness, speech recognition, bioelectrical impedance analysis (bia), face detection, system identification, elderly, back propagation artificial neural network (bp-ann), data analysis, infrared, galaxies
5	kl divergence, k-nearest neighbor algorithm, dependent bayesian probability models, embedded grammars, forming, qoe, minimum spanning tree, offloading, active shape model, computer-assisted diagnosis
6	smoothing l-1/2 regularization, multilingual, feedforward neural networks, bioinformatics, convergence, artificial neural network, functional linkage network, data distribution, complex-valued artificial neural network, coefficients modification
7	microphone array, knowledge base, expert systems, rasta filter, part of speech, computational drug repositioning, neural network, pid control, surface roughness, geiger-muller (g-m) detector
8	kalman filter, streptomyces avermitilis 41445 uv 45(m)3, confidence measures, web services, contrast-enhanced ultrasound, hmi, neural network, fault detection, emotion regulation, asymmetric membership function
9	risk assessment, fuzzy neural network, h-clat, sh test, major adverse cardiac events, knowledge base, fault diagnosis, potency, statistical parametric speech synthesis, softplus nonlinearity
10	knowledge base, agricultural machinery, back-propagation training optimization, autonomous landing, expert systems, self-organization, interface agents, traffic flow, correlation degree, magnetic resonance spectroscopy

Using the topic mapping algorithm, the evolutionary process of the topics, such as generation, integration, differentiation and extinction, could be derived. Fig. 6 shows the topic evolution of papers related to industry during three stages by using Sankey diagram. Each column is the topic distribution in that stage. During limited space, only the top three keywords with the highest probability are used to represent the topic. From the Sankey diagram, we can see the corresponding relationships between topics in different stages. The experimental results have revealed to some extent the topic evolution process of AI UIC research.

Table 6. Topic Distribution of AI UIC research in 2015-2017.

Id	Topics
1	machine learning, deep learning, neural network, artificial neural network, convolutional neural networks, deep neural network, speech recognition, face recognition, classification, automatic speech recognition
2	speech perception, neural networks, linear matrix inequalities, classification, cnn, road extraction, security, adaptive control, reconstruction, text analysis
3	complex network, neural network model, multiple coincidences, inter-crystal scatter (ics), gravitational lensing, feature enhancement, convolutional neural network (cnn), fuzzy logic, parkinson's disease, network speech recognition
4	asr, hidden markov model, developmental, global exponential stability, screening, speech synthesis, artificial neural networks, memristor-based neural networks, neurotoxicology, in vitro and alternatives
5	gmm-derived (gmmd) features, deep neural networks (dnn), cd-dnn-hmm,map, speaker adaptation, speaker adaptive training (sat), expert systems, uncertainty, jtc, human pose estimation, atr,temperature control
6	gabor wavelets, machine learning, rbf neural network, decision tree, rnn, bl lacertae objects, discriminant learning, high-value information discovery, c32, tree species
7	expert system, screening, distributed computing, neural network training algorithm, clinical decision support system, phone scams, spike transmission, pseudo-nitzschia, network speech recognition, electroencephalography (eeg)
8	autoencoders, detection, multi-objective optimization, artificial neural networks, principal component analysis (pca), aircraft air conditioning system, class variable or dependent variable, density based clustering algorithm, syncope, lesser cost
9	physicochemical exclusion rules, hot-spot temperature of windings, finite state transducer, decision analysis, distributed als recommender, cylinder pressure, recommendation, adaptation, narx, competition
10	artificial neural networks (anns), multi-agent systems, polyp detection, sensitivity analysis, patient handoff, genetic disorders, computer-aided detection, spiking neural network, auto disposition, lung cancer

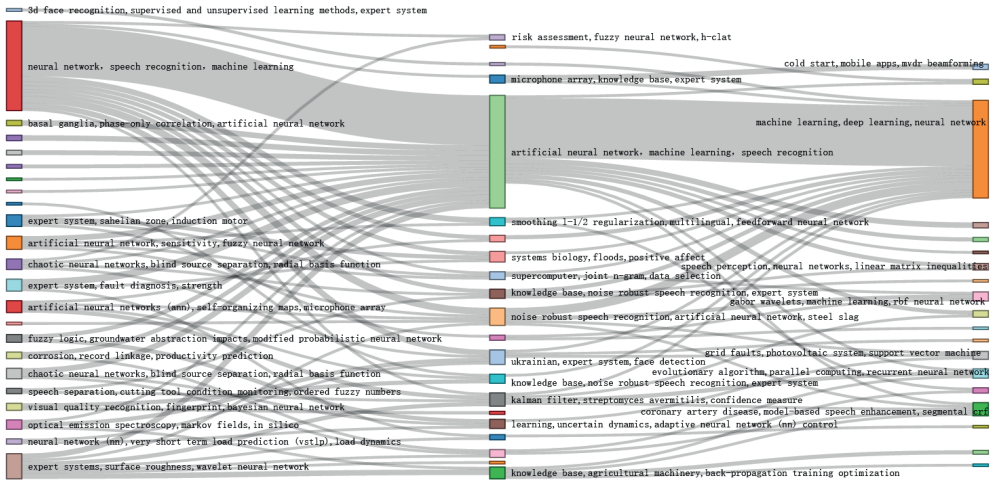


Figure 6. Topic evolution of AI papers related to industry during three stages (2008-2011, 2012-2014, 2015-2017).

Conclusions

In this paper, the open science behaviour of industry is studied in AI research field. The authors are classified into five types (*UI, UU, U, II, I*) according to the cross-institutional involvement. Cross-institutional authors have a relatively high proportion of all authors. Especially *UI* authors, they play an important bridge role in the UIC. Topic modelling method is used to reveal the underlying semantic topics within the papers. The results reveal the collaboration patterns and collaboration topics from the perspective of cross-institutional authors.

There are also some limitations and future directions of this study. Firstly, the motivation of the open science behaviour of industry is not explored. Why do industries voluntarily adopt open science strategies or behaviours? Why do scientists engage in both industry and university? Secondly, the relationship between the open science behaviour and innovation performance of industries is also a worthy research direction in the future work.

Acknowledgments

This work is partially supported by grant from National Natural Science Foundation of China (No. 71704019) and the Planning Fund for Liaoning Social Science (No. L17CGL009).

References

Åstebro, T., Bazzazian, N. & Braguinsky, S. (2012). Startups by recent university graduates and their faculty: Implications for university entrepreneurship policy. *Research Policy*, 41(4), 663-677.

Ankrah, S. N., Burgess, T. F., Grimshaw, P. & Shaw, N. E. (2013). Asking both university and industry actors about their engagement in knowledge transfer: What single-group studies of motives omit. *Technovation*, 33(2-3), 50-65.

Arel, I., Rose, D. C. & Karnowski, T. P. (2010). Deep machine learning-a new frontier in artificial intelligence research. *IEEE computational intelligence magazine*, 5(4), 13-18.

Azagra-Caro, Joaquín, M., Tijssen, R. J. W., Tur, E. M. & Yegros-Yegros, A. (2019). University-industry scientific production and the great recession. *Technological Forecasting and Social Change*, 139.

Callaert, J., Landoni, P., Van Looy, B. & Verganti, R. (2015). Scientific yield from collaboration with industry: The relevance of researchers' strategic approaches. *Research Policy*, 44(4), 990-998.

Chai, S. & Shih, W. (2016). Bridging science and technology through academic–industry partnerships. *Research Policy*, 45(1), 148-158.

Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M. & Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542(7639), 115.

Etzkowitz, H., Webster, A., Gebhardt, C. & Terra, B. R. C. (2000). The future of the university and the university of the future: evolution of ivory tower to entrepreneurial paradigm. *Research policy*, 29(2), 313-330.

Gulbrandsen, M. & Slipersaeter, S. (2007). The third mission and the entrepreneurial university model. *Universities and strategic knowledge creation*, 112-143.

Hicks, D. (1995). Published papers, tacit competencies and corporate management of the public private character of knowledge. *Industrial and Corporate Change*. 4(2): 401-424.

Hottenrott, H. & Lawson, C. (2017). A first look at multiple institutional affiliations: a study of authors in Germany, Japan and the UK. *Scientometrics*, 111(1), 285-295.

Jong, S. & Slavova, K. (2014). When publications lead to products: The open science conundrum in new product development. *Research Policy*, 43 (4): 645-654.

Leydesdorff, L. & Etzkowitz, H. (1996). Emergence of a Triple Helix of university-industry-government relations. *Science and public policy*, 23(5), 279-286.

Murray, F. (2002). Innovation as co-evolution of scientific and technological networks: exploring tissue engineering. *Research Policy*, 31(8/9):1389-1403.

Ponomariov, B. (2013). Government-sponsored university-industry collaboration and the production of nanotechnology patents in US universities. *The Journal of Technology Transfer*, 38(6), 749-767.

Brin, S. & Page, L. (1998). The Anatomy of a Large-Scale Hypertextual Web Search Engine, *Proceedings of the seventh International Conference on the World Wide Web* (WWW1998):107-117.

Simeth, M. & Lhuillery, S. (2015). How do firms develop capabilities for scientific disclosure? *Research Policy*, 44 (7): 1283-1295.

Teh, Y. W., Jordan, M. I., Beal, M. J. & Blei, D.M. (2006). Hierarchical dirichlet processes. *Journal of the American Statistical Association*, 101(476): 1566-1581.

Tijssen, R.J.W. (2012). Co-authored research publications and strategic analysis of public-private collaboration. *Research Evaluation*, 3: 204-215.

Wong, PK. & Singh, A. (2013). Do co-publications with industry lead to higher levels of university technology commercialization activity? *Scientometrics*, 2: 245-265.

Zhou, P., Tijssen, R. & Leydesdorff, L. (2016). University-industry collaboration in China and the USA: A bibliometric comparison. *PloS one*, 11(11), e0165277.

Globalization of Science: Evidence from authors in academic journals by country of origin*

Vít Macháček[1,2]

[1] *vit.machacek@cerge-ei.cz*
CERGE-EI, Politických vězňů 7, 110 00 Prague (Czechia)

[2] Institute of Economic Studies, Faculty of Social Sciences at the Charles University, Opletalova 26, 110 00 Prague (Czechia)

Abstract

The scientific community shares a common sense that by default, science should be globally oriented. This study measures the tendency to publish in globalized journals on a large dataset of journals indexed in the Scopus database. Based on data on 34 964 journals indexed in the Scopus Source List (Scopus 2018), we derived eight globalization indicators. These were subsequently scaled-up to the level of 174 countries and 27 disciplines between 2005 and 2017. The methodology draws from the pioneering work of Zitt and Bassecoulard (1998; 1999). The paper is accompanied by the interactive publication available at http://www.globalizationofscience.com

Advanced countries tend to have high globalization that is not varying across disciplines. Social sciences are similarly globalized as life sciences. The globalization in the former Soviet bloc is lower, especially in social sciences or health sciences. China has profoundly globalized its science system; gradually moving from the lowest globalization rates to the world average. Contrary Russia was constantly among the least globalized during the whole period, with no upward trend.

Introduction

The scientific community shares a common sense that by default, science should be globally oriented. Academic researchers present their results to their peers across the world to move the research frontiers beyond the existing boundaries. Doing so, they publish in journals contributed by researchers from the whole world.

This paper measures the globalization of science from the perspective of academic journals. On the most general level, the study asks whether researchers publish in journals presenting scientific results to the widest possible audience. The global dimension of the audience is stressed (therefore globalization), but also alternative specifications based on language and institutional concentration are added.

We explore the heterogeneity of globalization of science across countries and disciplines and in time. While the heterogeneity between disciplines can result naturally from different publication patterns of researchers, the cross-country heterogeneity within the same discipline points towards the research evaluation in the country and the research culture in a broader context. Systemic isolation of the country research output is indicative of the incentives provided by the evaluation system. Researchers in such country seem to lack sufficient motivation to publish in globalized journals. The development in time yields important information about the transformation of research in the country.

* Financial support from the research programme Strategy AV21 of the Czech Academy of Sciences and from the Charles University Grant Agency (GA UK; project no 1062119) is gratefully acknowledged. All usual caveats apply.

In Western Europe and North America, research has undergone a dramatic transition from the national to the transnational model of publications already in the 1980s and the beginning of the 1990s (Zitt et al. 1998). Thirty years later, the national journals still play an important role in the countries from the former Soviet bloc (see Moed 2018, Kirchik and Gingras 2012 or Figure 1).

This work follows Zitt and Bassecoulard (1998 and 1999), but since then any systematic evidence is missing. Still, the global research landscape changes. The world research production grows in the traditional research countries, but new research capacities are built from the ground up in the developing countries. The globalization of science adds an important insight into the transformation of research across the world.

Based on data on 34 964 journals indexed in the Scopus Source List (Scopus 2018), we derived 8 globalization indicators. These were subsequently scaled-up to the level of 174 countries and 27 disciplines between 2005 and 2017. The methodology allows for comparability between countries, disciplines, and in time. The main goal of the paper is to identify the most important global trends, but results allow for using the data with much higher granularity.

Globalization of science should not be confused with its quality (or relevance); they are likely to be related in many ways, depending on the discipline, but they are different phenomena.

In the first section, the existing literature is summarized. The second section describes data collection; the third describes the methodology and some of its limitations and the fourth results. The last section concludes. The paper is accompanied by an already released interactive study available at http://www.globalizationofscience.com (Macháček and Srholec 2019). Readers can spend their time with the interactive application, as it offers an intuitive way of exploring the results of this paper.

Globalization of science

The global research system grows both in terms of size (Royal Society 2011), and interconnectedness. New countries incorporate their research into global knowledge flows (Gazni, Sugimoto, Didegah 2012; Wagner et al. 2015). The international collaboration drives the growth of the research output (Adams 2013). Developing countries invest heavily to improve its research infrastructure (Royal Society 2011) and international visibility (Zhou and Glanzel 2010).

The globalization of science (from now on just globalization) contains a piece of useful information about this transformation process. The researchers in the country and discipline with high globalization publish their work in journals operating worldwide. On the contrary, when globalization is low, local researchers rely on journals that significantly deviate from the world distribution of journals. These journals are relevant for only a small group of local researchers.

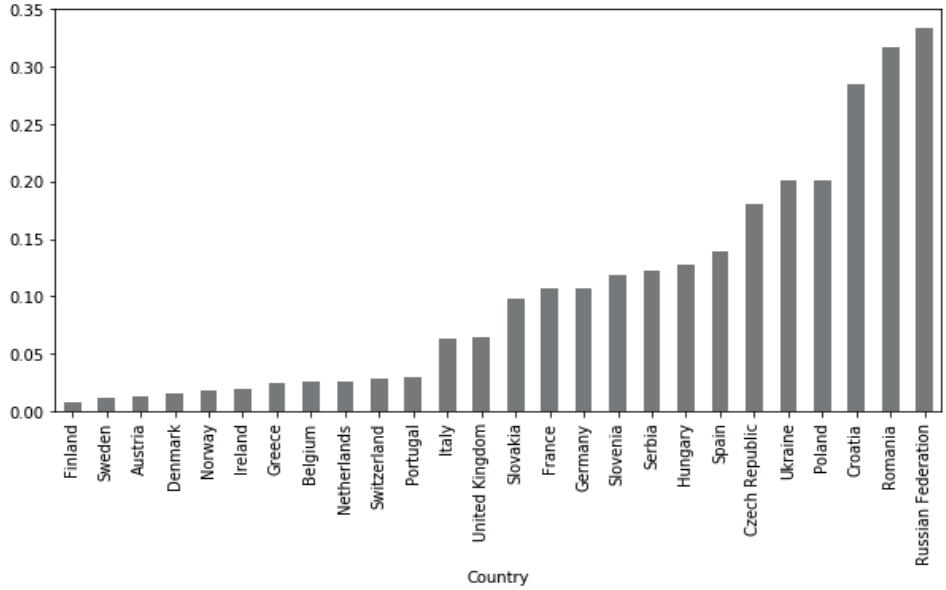
We perceive globalization in terms of journals. The academic journals are an essential platform for scientific knowledge dissemination. They allow scientists across the world to keep up-to-date with the latest discoveries and to present their results to the global audience. The naïve intuition suggests that the *globalized* journals are likely to be better serving their dissemination goal than journals operating in only a handful of countries (Buela-Casal et al. 2006).

Buela-Casal et al. (2006) warn against certain ambiguity connected with the term internationality. The definition is key to determine journal internationality. The journal can be considered international when contributed by researchers from more than one country or all countries in the world. Another definition considered journal international if its editorial board is international. The journal can be considered international if it publishes in English or even when it has a word "*international"* included in its name. It is also possible to analyze the internationality of citations of the journal. All these definitions often lead to different rankings of journal internationality. It is always necessary to clarify what is meant by internationality.

This key issue is tackled twofold: a) we analyze journals exclusively from the perspective of authors b) the possible ambiguity of results is overcame by employing a variety of indicators. The indicators emphasize country distribution but are also supplemented with indicators based on language and institutional concentration of authors. We refer to the globalization of science to emphasize the global reach of journals that is inherently present in half indicators.

After the fall of the Berlin Wall, the formerly separated research systems on both sides of the iron curtain started to merge. Simultaneously, the transition from the national to the international model of publication was happening at least on the western side of the curtain. The *transnational* English-written journals almost entirely replaced the *national* journals as a major publication platform (Zitt et al. 1998).

Figure 1: Share of research output flowing into domestic journals in Europe in 2015-2017



Note: domestic journals defined as journals with at least 33% authors from the same country as the publisher of the journal. Only articles, reviews, and conference papers are included. Publisher country from Scopus Source List is used to identify the domicile of the journal

Source: own calculation, Scopus; Scopus Source List

More than 25 years later a similar transition is still not finished in former Soviet bloc. Figure 1 shows that domestic journals are much more prevalent in these countries. For this figure, we define domestic journal as one having more than a third of authors from the same country as a publisher. More than 30 % of Russian research is published in Russian domestic journals. On the contrary, the same figure is around 1 % in Finland, Sweden or Austria. However, the still important role of national journals is not systematically addressed in the literature.

Domestic journals were originally published in national languages. However, they often adjust to the transforming environment by translating their content into English while keeping their author-base local (Kirchik and Gingras 2012, Moed 2018). Language of publication is not necessarily a good indication of the journal's globalization. Nevertheless, one indicator based on language is included in the paper.

Many journals operate worldwide and forcing a single-country domicile can be too strict. The concept of domestic journals outlined above also does not take into account the size of the research sector in the country. The composition of the editorial board can be a useful measure of journal internationality, but its data availability is limited. Moreover, it is easier to fake than the authors contributing to the journal. The journal's authors provide a direct measure of the journal globalization that is both available and relatively reliable.

The globalization is informative about the incentives provided by the research evaluation system in the country and the power relations in the local research system. Incentives influence researchers' decision on where to publish (Franzoni et al. 2011). The decision where to submit a paper is a key stage of the research-production process, which involves a highly strategic behavior of the researcher (see Heintzelman and Nocetti, 2009) His decision is linked to the strength of the "adverse mechanisms" - persistent proximity networks hindering the internationalization process (Zitt and Bassecoulard 2004). The local ties between researchers, institutions, research managers, and policy-makers can diverge the knowledge flows from international, to more local paths.

To our knowledge, the only paper providing analyzing globalization in countries is Zitt and Bassecoulard (1999). Based on journal-level indicators suggested in their previous paper (1998), they analyzed countries and discipline according to their internationalization. They acknowledge a "general trend towards internationalization," with some exceptions such as Russia. Since then, any systematic evidence is missing.

Data

The analysis is based on Scopus data. Scopus indexes approximately twice more journals than its main competitor, Web of Science (Mongeon and Paul-Hus 2016). It is more likely to contain the more localized part of the scientific output in the country.

The data for all 34 964 journals indexed in the Scopus Source List (Scopus 2018) were downloaded using Scopus API in August 2018. For each journal in each year between 2005 – 2017, we downloaded the country and institutional distribution of authors and the distribution of languages. Data were limited to articles, reviews, and conference papers.

The Scopus Search API was requested with the following query:

ISSN(AAAA-BBBB) AND DOCTYPE(AR OR RE OR CP) AND PUBYEAR = YYYY

in which AAAA-BBBB is the journal's ISSN and YYYY is the year. Rather than publication-level data, the aggregate distribution is collected. For each journal in each year, we collect the number of articles affiliated to each country, language, and institution.

Scopus Source List also contains Scopus Journal Classification (see Scopus 2019) used to assign journals to disciplines. Both more narrow classification (*Major Subject Classification;* referred to as *narrow disciplines*) and broad classification on 4 disciplines (*Broad Subject Clusters*) is used, supplemented by all-encompassing discipline *All* (*broad disciplines*).

Approximately 5 % of publications affiliated to the *undefined* country were excluded from the analysis. Undefined publications were also subtracted from the total number of publications in the journal. The data for Russia and the Russian Federation and Yugoslavia and Serbia were merged.

Globalization Indicators

The paper proposes eight *globalization indicators* assessing each country and discipline's globalization for each year. The two-step methodology builds on the pioneering work of Zitt and Bassecoulard (1998) and (1999). First, we calculate the globalization indicators for each academic journal indexed in the Scopus Source List in each year. Subsequently, the journal-level indicators are scaled up to the level of countries and disciplines.

Journal-level Indicators

The indicators evaluate each journal using a relatively simple metric. The variety of indicators increase the robustness of results. For most indicators, the main idea is that most globalized journals have a structure that closely resembles the global structure of the whole discipline. The likelihood to be published should not be affected by the origin of the researcher. The distribution of authors should correspond to the distribution of their respective disciplines.

The indicators of globalization vary in terms of input data (see Table 1). Half of the indicators take into account the whole distribution of the underlying data. Two indicators analyze only the three most important contributors in the journal. Two indicators are a simple share of documents fulfilling some condition. Six indicators analyze the country data. One employ language and one institutional data.

Half indicators compare the country distribution of authors to the distribution common in the discipline. The benchmark distribution is always calculated from all available periods so that the development in time takes into account the world trend. The rest does not consider discipline publication patterns.

$N_{c,j,y}$, $N_{c,d,y}$ and $N_{i,j,y}$ are the number of documents with authors affiliated to the country *c* or institution *i,* in journal *j* or discipline *d*, in year *y*.

$N_{LOCAL,j,y}$ is the number of documents with authors from the same country as the publisher of journal *j* in the year *y*. $N_{ENG,j,y}$ is the number of English-written documents in the journal *j* in year *y*.

$T_{j,y}$ denotes the total number of documents in the journal *j* in year *y*. Note that documents by authors from multiple countries are fully attributed to each country, i.e. $T_{j,y} \neq \sum_c N_{c,j,y}$.

The vectors $x_{c,j,y}$ and $m_{c.d}$ represent the country distribution of authors of the journal j and the discipline d, in which $x_{c,j,y} = \frac{N_{c,j,y}}{T_{j,y}}$ and $m_{c,d} = \frac{\sum_y N_{c,d,y}}{\sum_y \sum_c N_{c,d,y}}$. While $x_{c,j,y}$ is calculated separately in each year y, $m_{c.d}$ relates to the whole period.

Table 1: Globalization Indicators

Indicator	*Formula*	*Data*	*Bench.*	*Dist.*	*Source*
Euclidian distance (*euclid)*	$\sqrt{\sum \left(x_{c,j,y} - m_{c,d}\right)^2}$	Country	Yes	Full	Zitt and Bassecoulard (1998)
Cosine distance (*cosine*)	$\frac{\sum(x_{c,j,y} m_{c,d})}{\sum(x_{c,j,y}^2)\sum(m_{c,d}^2)}$	Country	Yes	Full	Zitt and Bassecoulard (1998)
GiniSimpson Index (*GiniSimpson)*	$1 - \sum \frac{N_{c,j,y}^2}{\left(\sum N_{c,j,y}\right)^2}$	Country	No	Full	Aman (2016)
Weighted Gini (*wGini*)		Country	Yes	Full	--
Largest Contributors Surplus (*top3*)	$\sum_{c=1}^{3} \left(x_{c,j,y} - m_{c,d}\right)$	Country	Yes	Partial	--
Institutional Diversity (*instTOP3*)	$\sum_{i=1}^{3} \left(N_{i,j,y}/T_{j,y}\right)$	Institutional	No	Partial	--
English Documents (*sEnglish*)	$\frac{N_{ENG,j,y}}{T_{j,y}}$	Language	No	Share	Buela-Casal et al (2006)
Local Authors (*localShare)*	$\frac{N_{LOCAL,j,y}}{T_{j,y}}$	Country	No	Share	Zitt and Bassecoulard (1998)

Aggregation

In the second stage, the journal-level indicators were aggregated to the level of countries and disciplines. The result $G_{c,d,y,i}$ is a weighted average of individual journals scaled between 0 and 1, where 0 is the lowest globalization across all years, countries and disciplines within the particular indicator and 1 is the highest.

To increase robustness and decrease volatility, the aggregation was only performed when the authors from the country published in at least 30 journals. This leads to gaps in results, particularly in the small disciplines and small countries.

The globalization of science in country c, discipline d and year y expressed by an indicator i is calculated as follows:

$$G_{c,d,y,i} = \sum_{j=1}^{J} a_{c,d,y,j} \cdot g_{j,d,y,i}(.)$$

$a_{c,d,y,j}$ is the share of documents with authors from country c in journal j on all documents from the country c, discipline d in year y, $g_{j,d,y,i}(.)$ is the globalization indicator i of journal j in the discipline d and year y.

Subsequently, the aggregated globalization index was standardized between 0 and 1 and converted to an ascending scale to simplify the interpretation of the results:

$$\mathrm{G}^{\mathrm{S}}_{\mathrm{c,d,y,i}} = \frac{\mathrm{G}_{\mathrm{c,d,y,i}} - \mathrm{G}^{\min}_{\mathrm{i}}}{\mathrm{G}^{\max}_{\mathrm{i}} - \mathrm{G}^{\min}_{\mathrm{i}}} \alpha_i$$

in which $\mathrm{G}^{\min}_{\mathrm{i}}$ and $\mathrm{G}^{\max}_{\mathrm{i}}$ is minimum and maximum value of the indicator i across all years, countries and disciplines and α_i equals -1 for the minimizing indicators (low values for high globalization) and 1 otherwise.

It is possible to compare globalizations between countries, discipline, and in time. However, the meaningful comparison between indicators is not possible due to the large heterogeneity of underlying distributions. The same value from two indicators cannot be interpreted as corresponding levels of globalization.

Limitations

The major drawback of the methodology is the representativeness of the underlying data. We refer to the *Globalization of Science*, but it might be more convenient to refer to the *Globalization of Science In Journals Then Indexed By Scopus*. Scopus indexes the research output across units unevenly. Some units are underrepresented in the dataset. The representativeness issue is present in all major dimensions – countries, disciplines, and in time.

It is well possible that Scopus contains a larger part of the content published in the Netherlands than that from India or other developing countries. We are not aware of any study analyzing major databases representativeness across countries. However, it is reasonable to assume that the Scopus database contains the more globalized part of the scientific production within a country. Therefore, the results show the upper bound of globalization of the possibly underrepresented countries.

As the publication behavior varies across disciplines, we use data for all three major document types in the journal articles – articles, reviews, and conference papers. However, publications outside of academic journals are neglected. In this sense, the representativeness can seriously impact the interpretation of disciplines strongly relying on other modes of publications, such as Arts and Humanities.

The results are sensitive to Scopus editorial decisions on indexing of titles. Large year-by-year jumps are not necessarily caused by fundamental changes of the researchers' behavior but are mostly driven by adding (or removing) journals in Scopus. For example, in 2007, Scopus indexed 184 documents from Romania in Social Sciences. One year later, this figure almost tripled to 471, while the globalization decreased from 0.5 to 0.2. This development is driven by a sudden indexation of several Romanian journals[1] that publish mainly Romanian researchers. Short-term changes have to be interpreted with caution.

[1] Namely Amfiteatru Economic Journal, Transylvanian Review of Administrative Sciences, Journal of Economic Computation and Economic Cybernetics Studies and Research etc.

The disciplines are still relatively broadly defined. Each discipline can contain many diverse research topics with varying publishing patterns. If these are strongly antagonistic, the aggregate globalization can be distorted.

Globalization can be conflicted with the quality of the research. The strong concentration of high-quality research in only a handful of countries under certain conditions leads to lower globalization just because doing quality research. However, the countries where research is commonly considered as a high quality usually have high globalization. The impact of this issue on results is very limited.

Results

The reported results $G^{S}_{c,d,y,i}$ are normalized relative to all observations within a single indicator. $G^{S}_{c,d,y,i} = 0$ and 1 always refer to the least and most globalized country, discipline and the period within all results available within a single indicator. The methodology allows for any comparison within an indicator. However, the comparison between indicators is not possible as the journal-level indicators distributions vary.

After excluding all dependent territories except Hong Kong, the computation algorithm yielded results for 174 countries. The minimum requirement on the number of journals to calculate results lead to gaps in the data. Naturally, the broader the discipline definition and the larger the research production, the more countries are available. In 2017, the data were available for 171 countries in the discipline *All,* but for most of the narrow disciplines, the results are calculated for less than 100 countries. The results availability also grows in time together with output indexed in Scopus.

The results are robust to varying indicators. The correlation coefficients reported in table 2 are generally high. 25 out of 28 coefficients exceed 0.5, and a half coefficients are higher than 0.7. Also, visual check in the interactive application shows that most globalization paths in time and relative rankings are similar across indicators.

The most in-between indicator is Euclidian distance with a correlation coefficient higher than 0.75 in 8 out of 9 indicators. Therefore Euclidian distance is used as a main indicator of globalization. By default, we refer to it when not stated otherwise.

Table 2: Full-Sample Correlation Matrix of Indicators

Indicator	*euclid*	*cosine*	*GiniSimpson*	*wGini*	*top3*	*instTOP3*	*sEnglish*	*localShare*
euclid	1.00	.83	.87	.90	.93	.81	.61	.75
cosine	.83	1.00	.64	.93	.75	.69	.47	.41
GiniSimpson	.87	.64	1.00	.79	.72	.67	.64	.78
wGini	.90	.93	.79	1.00	.86	.73	.55	.56
top3	.93	.75	.72	.86	1.00	.79	.51	.67
instTOP3	.81	.69	.67	.73	.79	1.00	.43	.57
sEnglish	.61	.47	.64	.55	.51	.43	1.00	.61
localShare	.75	.41	.78	.56	.67	.57	.61	1.00

Pearson correlation coefficients of all available data for each indicator; Source: Scopus; own calculation

In what follows, only the major trends of globalization are summarized. The data are available in much greater detail in the interactive application, where the readers are invited to explore the situation in country, discipline, or period of their interest. The data are also available in Appendix 1.

The results cover the period between 2005 and 2017. When analyzing cross-country differences, the period 2015 – 2017 is considered. The three-year window allows to partially balance the effect of jumps caused by changes in the Scopus indexation.

The narrow disciplines are neglected. The data availability in smaller countries is lower than in broader disciplines, and the identification of general trends makes more sense in the broader context. Generally, the narrow disciplines usually follow similar paths as the broad disciplines. However, the volatility naturally grows with detail.

Countries were assigned to country groups according to their economic status by IMF (2003) classification. The countries are divided into three categories: (a) *Advanced countries* cover the richest countries in the world in the mainly in Western Europe, North America, and Eastern Asia. (b) *Transition countries* consist mainly of the post-soviet countries in Central and Eastern Europe, Central Asia, and Cuba. (c) *Developing countries* – the rest of the World, including China. See Appendix 2 to see exact classification.

The research is most globalized in *advanced countries*, followed by *developing countries* and *transition countries*. Table 3 presents the means and standard deviations of all globalizations within country groups between 2015 and 2017. In all disciplines, the mean is highest in *advanced countries* and the lowest in *transition countries*. The *life sciences* and *physical sciences* are more globalized than *social sciences* and *health sciences*.

Table 3: Means and standard deviations of globalizations within country groups in 2015-2017 measured by Euclidian distance by disciplines (wider definition)

Economic status	All	Life	Physical	Social	Health
Advanced countries	.75	.80	.79	.75	.72
	(.03)	(.03)	(.03)	(.05)	(.04)
Developing countries	.70	.74	.75	.68	.68
	(.06)	(.07)	(.05)	(.10)	(.08)
Transition countries	.61	.68	.66	.53	.58
	(.09)	(.10)	(.11)	(.13)	(.13)

Source: own calculation; IMF (2003); Scopus; Standard deviations in brackets.

The science in *advanced countries* is not only highly globalized but also relatively invariant. The figures are similar in all disciplines, and their standard deviations are low. In all disciplines, research in advanced countries tends to be globalized. Only research in Japan and South Korea is somewhat lower (globalization 0.66 in discipline *All*). In Germany, France, and Spain, the *life sciences* and *physical sciences* are highly globalized (approx. 0.78), but *social sciences* and *health sciences* tend to be below standards (<0.65).

In *transition countries,* the globalization is generally lower in all disciplines. The means vary between 0.53 in Social Sciences, to 0.68 in Life Sciences. The variance is also larger within disciplines. The standard deviations are approximately three times larger than in advanced

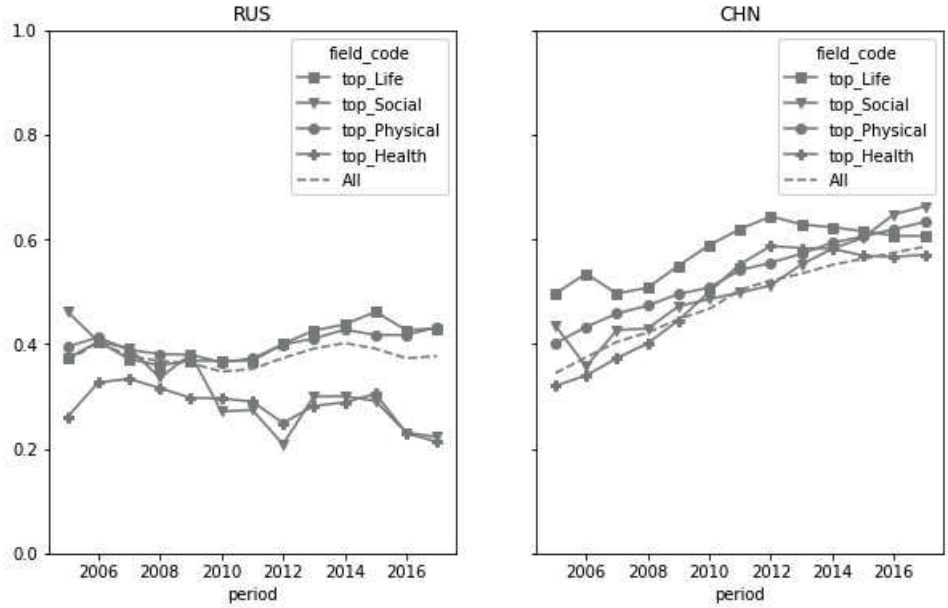
countries. It is hard to characterize this diverse group. However, there is only one country whose characteristics fits well with the patterns common in advanced countries – Estonia.

In some countries, the globalization is relatively high in *life sciences* and *physical sciences* (>0.7), but significantly lower in *social sciences* and *health sciences* (<0.55). This applies mainly countries from the former Soviet bloc in Europe, such as Czechia, Poland, Slovakia, Slovenia, or Croatia. However, there are also countries who perform badly in most disciplines – Belarus, Ukraine, Kazakhstan, or Russia.

Russia is a prime example of the strongly isolated research system (see Figure 2). In 2017, Russia ranked as the first or second least globalized country in all broad disciplines. Even when extended to the narrow definition of disciplines, Russia is among the last 3 countries in 23 out of 25 disciplines where the data are available.

Where there is a will, there is a way. At the beginning of the analyzed period, in 2005, China was the least globalized country in the world in *All disciplines,* and among the 5 least globalized in all broad disciplines (see Moed 2002). China invests heavily in the modernization of its science infrastructure (Royal Society 2011). Figure 2 shows a rapid transformation of the Chinese system that resulted in the relatively fast growth of globalization. During the analyzed period, the globalization grew fast across all disciplines. In 2017 China was still less globalized than *advanced countries*, but its progress is visible.

Figure 2: Globalization in broad disciplines in China and Russia in time (Euclidian distance)



Source: own calculation, Scopus

The map on Figure 3 shows that in all BRIICS[2] countries, the globalization is below (or very close to at maximum) average in all broad disciplines in 2017. The long-term growth trend in the past 12 years is only seen in China. In Indonesia, the globalization even slightly decreases in time in all broad disciplines. The path to globalization is far from being guaranteed for the developing countries.

Figure 3: Globalization of science on the world map (2017; Euclidian distance; All disciplines)

Source: own calculation; Scopus; the darker the color, the higher globalization

Conclusions

The paper measured the globalization of science in 174 countries and 31 disciplines between 2005 and 2017. Using the data on articles, reviews and conference papers from journals indexed in the Scopus Source List we developed a two-step methodology that scales the globalization of individual journals up to the level of countries, disciplines and years. Multiple indicators of globalization (based on country data, language data, and institutional data) were constructed to verify the robustness of our findings. The paper is also accompanied by an interactive application available at http://www.globalizationofscience.com/.

The transition from national to the transnational mode of communication took place in the 1980s and 1990s in Western countries. A similar transition is still not finished in the countries of the former Soviet bloc. Especially in the Social Sciences and the Health Sciences, the role of non-globalized journals is high. An example of Russia shows that the lack of globalization can be very persistent. An example of China shows that under certain conditions, the widespread integration into a global research publication flows is possible, and it can even be relatively fast.

The granular data on the globalization of science can serve as a good starting point for further exploration for researchers, research managers, policy-makers, and other research evaluation practitioners. These professionals should be asking whether the level of globalization in their context is adequate. However, further research necessary to understand the role of globalized journals in modern science. The relationship between quality and globalization is far from straightforward.

References

Adams, J. (2013). Collaborations: The fourth age of research. Nature, 497(7451), 557.

Aman, V. (2016). Measuring internationality without bias against the periphery. In 21st International Conference on Science and Technology Indicators-STI 2016. Book of Proceedings.

[2] Brazil, Russia, India, Indonesia, China and South Africa

Buela-Casal, G., Perakakis, P., Taylor, M., & Checa, P. (2006). Measuring internationality: Reflections and perspectives on academic journals. Scientometrics, 67(1), 45-65.

Franzoni, C., Scellato, G., & Stephan, P. (2011). Changing incentives to publish. Science, 333(6043), 702-703.

Gazni, A., Sugimoto, C. R., & Didegah, F. (2012). Mapping world scientific collaboration: Authors, institutions, and countries. Journal of the American Society for Information Science and Technology, 63(2), 323-335.

Heintzelman, M., & Nocetti, D. (2009). Where should we submit our manuscript? An analysis of journal submission strategies. The BE Journal of Economic Analysis & Policy, 9(1).

IMF (2003) World Economic Outlook (Statistical Appendix; pp. 163-169). Available at: https://www.imf.org/external/pubs/ft/weo/2003/02/

Kirchik, O., Gingras, Y., & Larivière, V. (2012). Changes in publication languages and citation practices and their effect on the scientific impact of Russian science (1993–2010). Journal of the American Society for Information Science and Technology, 63(7), 1411-1419.

Macháček, V. and Srholec, M. (2019) Globalization of Science: Evidence from Authors in Academic Journals by Country of Origin. 6/2019. Institute for Democracy and Economic Analysis (IDEA).

Moed, H. (2002). Measuring China" s research performance using the Science Citation Index. Scientometrics, 53(3), 281-296.

Moed, H. F., Markusova, V., & Akoev, M. (2018). Trends in Russian research output indexed in Scopus and Web of Science. Scientometrics, 116(2), 1153-1180.

Mongeon, P., & Paul-Hus, A. (2016). The journal coverage of Web of Science and Scopus: a comparative analysis. Scientometrics, 106(1), 213-228.

Scopus (2019) What is the complete list of Scopus Subject Areas and All Science Journal Classification Codes (ASJC)? Available at: https://service.elsevier.com/app/answers/detail/a_id/15181/supporthub/scopus/related/1/ [accessed May 28, 2019]

Scopus (2018) Scopus Source List (May 2018 version). Available at: https://www.scopus.com/sources

Wagner, C. S., Park, H. W., & Leydesdorff, L. (2015). The continuing growth of global cooperation networks in research: A conundrum for national governments. PLoS One, 10(7), e0131816.

Wilsdon, J. (2011). Knowledge, networks and nations: Global scientific collaboration in the 21st century.

Zitt, M., & Bassecoulard, E. (1998). Internationalization of scientific journals: a measurement based on publication and citation scope. Scientometrics, 41(1-2), 255-271.

Zitt, M., & Bassecoulard, E. (1999). Internationalization of communication a view on the evolution of scientific journals. Scientometrics, 46(3), 669-685.

Zitt, M., & Bassecoulard, E. (2004). Internationalisation in science in the prism of bibliometric indicators. In Handbook of quantitative science and technology research (pp. 407-436). Springer, Dordrecht.

Zitt, M., Perrot, F., & Barré, R. (1998). The transition from “national” to “transnational” model and related measures of countries' performance. Journal of the American Society for Information Science, 49(1), 30-42.

Zhou, P., & Glänzel, W. (2010). In-depth analysis on China’s international cooperation in science. Scientometrics, 82(3), 597-612.

Appendix 1 and 2 – Complete results available in CSV files data.csv and countries.csv. The files are available at: https://github.com/vitekzkytek/GlobalizationPaper/blob/master/appendix/

Predatory publications in Scopus: Evidence on cross-country differences*

Vít Macháček[1,2] and Martin Srholec[1]

[1] *vit.machacek@cerge-ei.cz; martin.srholec@cerge-ei.cz*
CERGE-EI, Politických vězňů 7, 110 00 Prague (Czechia)

[2] Institute of Economic Studies, Faculty of Social Sciences at the Charles University, Opletalova 26, 110 00 Prague (Czechia)

Abstract

The paper maps the infiltration of so-called "predatory" scholarly journals into the citation database Scopus. Using the names of "potential, possible, or probable" predatory journals and publishers on Beall's lists, we derived ISSNs of the respective journals from Ulrichsweb and searched Scopus with it. A total of 324 matched journals with 164 thousand documents indexed in Scopus over 2015-2017, making up a share of 2.8 % of the total articles have been identified. An analysis of cross-country differences in the tendency to publish in these journals shows that overall the most affected are middle-income countries in Asia and North Africa, however, this is a truly global problem striking across continents, cultures and political systems.

Introduction

The business model of so-called "predatory" scholarly journals is based on a paid open-access (OA) publication model: the publisher does not charge for subscription, but receives money directly from the authors for publication of article. As a result, the content is accessible for free to anyone. However, the predatory practice also entails a conflict of interests that has a potential to undermine the credibility of scholarly publishing. Authors are motivated to pay to have their work published for the sake of evaluation and career progression. In return, fraudulent publishers turn a blind eye to limitations of the submitted papers during peer-review. Predators' primary goal is to generate income from authors' fees. The worst of them fake peer-review and print anything for money, without scruples.

Why do researchers participate and offer their publications to predatory journals? Some of the predatory publication can be attributed to the low experience of young researchers (Xia et al 2015). They send their texts in a good faith that their text is going to be properly processed, but get cheated by the fraudulent journals. But some authors can also be trying to actively boost their career prospects. They send manuscripts to predatory journals to make their scientific results look better and are willing to pay the price of participating in the fraudulent scheme. If the local research environment accepts such results as a piece of solid scientific work and can help researcher to climb the ladder in the hierarchy, the motivation to pay for publishing the pseudo-scientific results grows.

Predatory publishing can be seen as wasting of resources. Bjork and Shen (2015) estimate the size of predatory market to 74 million USD in 2014 and the figure might have grown since. But these are only direct costs associated with Article Processing Costs (APCs). Perhaps more important than direct costs are the indirect opportunity costs. The opportunity to flaw the standard peer-review process distracts the concentration of researchers. Instead of fully

* Financial support from the research programme Strategy AV21 of the Czech Academy of Sciences is gratefully acknowledged. Earlier version of the paper was presented at the IDEA think tank seminar on "Predatory journals in Scopus: New evidence in a worldwide comparison", 16 November 2016, Prague, Czech Republic. The paper has benefited from comments and suggestions from Daniel Münich and participants at this event. All usual caveats apply.

focusing on generating the internationally competitive science, the researchers are motivated to write bogus pseudo-science that only pretends scientific relevance. If this happens on a massive scale on the national level the efficiency of the local research evaluation system is questioned. The fact that research is often funded from the public sources only amplifies the relevance of these concerns.

This study aims to comprehensively and systematically map the penetration of predatory journals in the Scopus citation database and present representative evidence on cross-country differences in the propensity of scholars to publish therein. The cross-country differences can be useful for understanding the phenomenon of predatory publishing itself and for delivering effective means to fight it. This is the first time representative evidence of such scale has been compiled. The results suggest which countries and regions should concern about the effectivity of their research systems and consider reforms enforcing publications comparable with international standards.

The first section explains the phenomenon of predatory publishing and reviews the existing literature studying the content of Beall's lists. The second section then explains how the dataset has been constructed and elaborates on limitations of the data. The third section provides an exploratory analysis of the cross-country differences and of geographical distribution of tendency to publish in predatory journals. The concluding section summarizes the key findings and pulls the strands together.

Predatory journals, Beall's lists and existing evidence

The term predatory publishing was coined by Jeffrey Beall on his blog (Beall 2016). It describes a practice of abusing paid OA scheme of scientific publishing. In difference to standard subscription-based models the authors publishing via paid OA pay Article Processing Costs (APC) directly to the journal publisher. Hence both the author and the publisher are financially motivated in publication of articles. Predatory journals are such that perform only vague, formal or even none peer-review and thus allow for publication of pseudo- or even non-scientific results pretending to be scientific (Bohannon 2013, Butler 2013). Other attributes of predatory journals such as aggressive marketing practices, fake members of editorial boards or amateur business management are often mentioned. But these are only side-effects of core defining elements. We use the term predatory journals to mark journals that are suspect of abusing paid open-access to attract fees from authors with significantly flawed editorial practice that fails in enforcing the minimum scientific quality requirements.

The OA model, while being one of defining elements of predatory journals, is not at fault per se. The inherent conflict of interest does not necessarily have to be exploited. The world research infrastructure already offers effective means to ensure the quality of editorial practice. Databases dedicated to support open-access such as DOAJ are already trying to develop operational mechanisms guaranteeing quality and employing transparency measures such as open peer-review can easily detect the fraudulent publishers. Journals not performing peer-review would have nothing to reveal. The existence of fraudulent journals does not mean that the movement calling for democratizing communication of science should be abandoned.

It is important to note that the predatory journals are not the exhaustive way the bogus science is published. There is no guarantee that the journal based on standard subscription model will be of a high quality. Also in these journals the editorial practice can be flawed and corrupted for example by exploiting personal links. However, the direct financial interaction between the journal and authors allows the OA journals to bend editorial practice more easily than in the

standard subscription. Moreover, up to our knowledge, there is no list of bogus subscription-based journals available for exploring.

We believe that the tendency to publish in predatory journals is related to the quality of research evaluation in the country. In countries where the culture of evaluation forces researchers to publish in respectable journals, the motivation to publish in predatory journals is rather low if any. Publication in dubious predatory journal can harm the researcher's reputation instead of boosting it. The more the research evaluation system relies on simple rules such counting of publications indexed in databases such as Scopus, Web of Science or Medline, the higher incentive for researchers to publish in bogus journals. The more the researchers work is evaluated by peer-review based system on all individual, institutional, grant funding and national level, the smaller the space for abusing it.

Harder than to define the predatory journal is to recognize it in practice. There is no clear boundary between journals that take their editorial practice seriously and those that are just a vehicle for exploiting publication fees. Most often black-lists are used to identify suspect predatory journals. The prime example is already mentioned Jeffrey Beall's blog (Beall 2016). He maintained two regularly updated lists of "potential, possible, or probable" predatory journals and publishers, henceforth for the sake of brevity referred to as "predators" only: i) a "*list of standalone journals*", which contains individual journals suspected of predatory practices that are likely to exist independently of any publishing house; and ii) a *"list of publishers*", which highlights questionable publishers, most of which print multiple journals. Beall's blog went off-line on January 15th, 2017 (Strausheim 2017). New black-list is currently offered by a private company Cabell's, but its high costs do not allow for in-depth study. China reportedly prepare their own "black-list of poor quality journals", however it is not yet available (Cyranoski 2018).

The inclusion of individual journals into the black-list should be based on more or less rigid criteria. Beall revealed the list of criteria that he admittedly uses for decision over journals and publishers (Beall 2015). Also Eriksson and Helgesson (2017) suggest a list of characteristics to identify predatory journals. We will concentrate on Beall's criteria as his list is used in our analysis. One set of indicators point directly to credibility of editorial practice: ("*Evidence exists showing that the publisher does not really conduct a bona fide peer-review"; "No academic information is provided regarding the editor, editorial staff, and/or review board members"*). Other group of indicators are indirect as they do not concern editorial policy, but other measures of professionalism and/or compliance with ethical standards ("*The publisher has poorly maintained websites, including dead links, prominent misspellings and grammatical errors on the website"*; "*Use boastful language claiming to be a 'leading publisher' even though the publisher may only be a start-up or a novice organization"*).

Crawford (2014a) investigated the content of Beall's list. The author went through every single item on Beall's lists (in late March and early April 2014). He found 9,219 journals in total, of which 320 were from the list of standalone journals and 8,899 from the list of publishers. Between 2012 and 2014 almost 40 % of those journals published fewer than four articles or none at all, in other words were empty shells, and that a further 20 % published only a handful of articles. Another 4 % consisted of dying or dormant journals with a quick drop to few articles in 2014, and 6 % were unreachable (the web link was broken, for instance). Hence only approximately one third of the identified journals published articles on a regular basis.

The lack of trustworthiness of Beall's listed journals and gaps in their editorial practice has also been demonstrated in Shamseer et al. (2017). They systematically compared the predatory journals with "ordinary" OA journals and also subscription-based journals. They proved that predatory journals much more often than other contained spelling errors, but also promoted bogus bibliometric metrics on their website. The authors also had much more often trouble to verify the connection of editorial board members with the journals. Famous field experiment by Bohannon (2013) proved the flawed editorial practice by submitting fake scientific articles to 304 predatory journals. More than half of them actually accepted the fake article for publication. Majority of Beall's listed journals raise very serious doubts about quality of their editorial practice.

To the best of our knowledge, there is only Shen and Bjork (2015) has attempted to estimate the regional distribution of predatory journals as a whole so far. On 1st September 2014 the authors manually identified 11,873 journals in total, from which they selected a stratified sample of 613 titles. Analysis of detailed data for this sample showed that three quarters of authors originated from Asia and Africa. As they themselves point out, however, these are only rough estimates.

Xia et al (2015) examined author profiles of 68 journals from Beall's list in the field of biomedical science, including data on the geographic distribution of nearly thousand authors. The analysis showed that the authors are based mainly in South Asia and Africa and that they are strongly concentrated in a few countries only, predominantly India but also Nigeria and Pakistan. The authors conclude that economic and sociocultural conditions prevailing in these developing countries contributed to the high inclination to publish in predatory journals. Nevertheless, as the authors also acknowledged, the sample size of the analysis is relatively small, which limits the findings at the country level and its focused exclusively on one field. More research on larger datasets is clearly needed to establish which countries are in the harm's way.

Neither research systems in developed countries are resistant to predatory publishing. Bugues et al (2017) showed that 5 % of researchers in Italian academia have published in journals included on the Beall's list. This study also mentions large incentive to publish in such dubious journals – the indexation in respected databases such as Scopus, that "many institutions consider a guarantee of quality" (p. 1). The study also indicate that journals from Beall's list tend to have low academic impact and cites Italian researchers admitting that editorial practice in these journals is truly flawed.

Some journals are surely more "predatory" than others. Some are truly fraudulent, while many others may be on the margin. But the form of Beall's list force us to work with a binary classification only, in which a journal is labelled either predatory or not. As Beall did not provide explanation of his decisions in a systematic manner, it is not possible to make any quantification of "predatoriness" of the journal, despite the fact that elaborated criteria existed. For this study, the journal either is listed by Beall or it is not.

A caution is warranted when working with Beall's list of publishers. Classifying the entire publishing house as predatory is a strong judgment, as it cannot be ruled out that alongside with truly fraudulent journals have been blacklisted also those that are fine. The list includes publishers maintaining broad portfolios of dozens and even hundreds of journals, so it might be that some of them strictly speaking do not deserve the predatory label and that using this list could result in an overestimation of the true "predators". It is likely that the vast majority of

these journals are of poor quality, but this is not a crime per se. After all, many non-predatory but local journals which publish research of marginal relevance, have also found their way into the main citation databases. One must therefore keep in mind that the list of publishers is a relatively rough brush.
The strong criticism of Beall's list emphasize the low transparency of the decision-making process (see Berger and Cirasella 2015, Crawford 2014b). Although the criteria are public, the justification of the decision over individual journals and publishers is often not clear and hardly verifiable. While on some important cases Beall published at least few sentences on his blog or on Twitter, very often the journal or publisher is only added to list, without any reasoning. The lack of comprehensive, rigid and formal justification of the Beall's judgement is still a major drawback of his list.

The largest controversies were triggered by inclusion of Frontiers Research Foundation in October 2015 on the list of publishers. According to critics of this decision the Frontiers publisher is "legitimate and reputable and does offer proper peer-review" (Bloudoff-Indelicato 2015, p.1). Clearly, Frontiers journals are quite far from typical predatory outlet without any relevance for scientific community. Actually, the citation rates ranks them fairly well. SJR index reveal that Frontiers is much more cited in comparison with other journals listed by Beall's. Most Frontiers journals are ranked in the first quartile (Q1) in at least one field according to Scimago SJR. Only 4 journals in Frontiers portfolio of 29 evaluated did not get into highest quartile. 3 of those ranked in Q2. Only one journal is ranked in last quartile only. Most of their journals are also indexed in Web of Science and Directory of Open Access Journals. Beall on the other hand defended his decision by pointing out several articles that according to him should not have been published.

Another concern arises from the timescale. The predatory status is derived from the content of Beall's lists on 1st April 2016. Jeffrey Beall continuously updated his lists. But the lists always reflect only the current status, with no indication of when the journal or publisher became predatory. From this follows that when looking back in time we may run into the problem of including in the predatory category records that actually do not deserve that label, because the journal switched to the predatory regime only recently. In other words, it well might be that older articles in journals that are currently considered to be predatory may in fact have gone through a standard peer-review. Hence, historical data must be used with caution.

Jeffrey Beall is American and his lists are probably subject to English bias. The lists contain mainly bogus journals that at least have their websites in English. In regions where large part of scientific output is written in other languages - such as in Latin America, China or in Francophone world - the estimates of the extent of predatory publishing based on Beall's list might be underestimated, since Beall did not identify predatory journals in local languages. Since majority of internationally relevant science is published in English this does not constitute a severe problem for our estimation strategy. However this bias has to be kept in mind when interpreting cross-country differences.

Scopus, rather than Web of Science, is used to construct the database, because it covers substantially more journals (Mongeon and Paul-Hus 2016); thus it is likely to be more susceptible to predators. Yet journals indexed in Scopus should fulfil minimum quality criteria, which among other things include performing a peer-review (Scopus 2016). Scopus-listed journals are also considered to be trustworthy by research evaluation and funding systems in many countries, such as, for example, the "coffee grinder" in the Czech Republic (Good et al (2015) or in Italy Bugues et al (2017). If a predatory journal becomes indexed in Scopus, the

motivation to publish in it gets a clear boost for scholars. The journals indexed in Scopus might represent only a tip of the iceberg of predatory publishing, but it is the most dangerous part as they are successful in pretending scientific quality.

The literature suggests that only minor part of all the listed journals is expected to be indexed in Scopus, but yet it might be quite substantial number of journals. The authors should most often originate from developing countries, especially from the countries with strong influence of English culture (due to language bias). The list of standalone journals is expected to be more reliable than list of publishers, but yet the list of publishers should follow similar patterns. We also expect that the tendency to publish is somewhat related to the culture of research evaluation in the country.

Database

The exhaustive database of Beall's listed journals indexed in the Scopus database has been constructed. First, by matching the two lists by Jeffrey Beall with records in the Ulrichsweb database (Ulrichsweb 2016), we built a comprehensive overview of journals suspected of predatory practices. The International Standard Serial Numbers (ISSNs) of these journals were then searched in Scopus and their bibliometric data were downloaded. As such not only a full list of predatory journals enlisted in Scopus, but also unique data on the composition of authors publishing in these journals by the country of origin.

Beall's lists were downloaded on 1st April 2016. First all search terms were identified in each item on the lists. For some entries Beall presented more versions of its name, for example the journal name and its abbreviation. All available versions were used as a search term. The identified search terms were searched for in the Ulrichsweb database on the same day, using an automatic script programmed in Python, while keeping the information about the item on Beall's list that is searched for. When searched for standalone journal the '*title'* field was used by the script. When searched for publisher, the '*publisher*' field was used. The search term was always enclosed in quotes. In the end the algorithm saved all search results. The search request in Ulrichsweb looked as follows for standalone journals:

+(+title:("Academic Exchange Quarterly")) or *+(+publisher:("Abhinav"))*

Ulrichsweb search engine uses 'fuzzy' search and does not require perfect matching of strings. For example when searched for publisher *Academe Research Journals*, also journals of *Academic Research Journals* were found. This is beneficial, because the search is robust to typos, interpunction signs and small errors written in the search terms. However it also require careful manual verification of the search results.

The raw search on Ulrichsweb resulted in the database of 19 141 search results linked to individual entries on the Beall's list. In next step the results without ISSN were removed as they are most probably not listed in Scopus neither. 16 037 search results were left after removal, but those are only 7 568 unique ISSNs. This is due to fuzziness of the Ulrichsweb search mentioned above and the searching for multiple search terms related to the same entry.

Further filtration was needed to exactly identify the journals truly listed by Beall. In order to resolve any inconsistencies, for example due to inaccurate names, the remaining search results were checked manually. Beall's lists consist of hypertext links. To check our entries the ISSN we compared the ISSN on the journal's website with the ISSN found on Ulrichsweb. If the two ISSNs matched, the entry was retained; if they differed, the entry was removed from our

database. Publisher's identity was confirmed if at least one ISSN listed on its web site was found in an entry linked to the publisher's name on Ulrichsweb. If a matching ISSN was found, all other journals of the same publisher were also retained in our database. If no matching ISSN was found, and the entries therefore referred to a different publisher, they were removed.

In total we confirmed 4 665 ISSNs associated with Beall's lists. Many journals have dual ISSNs, one for a printed and one for an electronic version. The number of individual journals is 3 295, of which 310 feature on the list of standalone journals and 2 954 refer to his list of publishers. Additional 31 journals were both in the list standalone journals and in the list of publishers. For simplicity we refer to journals from both lists as the journals from list of publishers only from now on.

This is roughly in line with the previous analysis of Crawford (2014a). There are 1 003 hypertext links on the list of standalone journals, from which follows that more than two thirds of them are not covered in Ulrichsweb, let alone in more selective databases, and so apart from the unverified information on their own web pages, there is no information about them. It is important to realize that there are many predatory journals not registered in existing databases.

Box 1. The Data Generation Process

1) Obtaining the "predatory" ISSNs
    a. Beall's lists downloaded on the 1st April 2016.
    b. The names on Beall's lists were searched for using an automatic script in Ulrichsweb on the same day.
    c. The entries found in Ulrichsweb were manually verified with the help of hypertext links in Beall's lists.
    d. 4 665 ISSNs of 3 295 individual journals were confirmed to be associated with Beall's lists.
2) Searching for ISSNs in Scopus
    a. The "predatory" ISSNs were searched for using an automatic script in Scopus in March 2018; the script downloaded the number of indexed documents and their distribution by author's country of origin.
    b. 439 ISSNs of 324 individual journals with at least one entry in Scopus were identified.
    c. To avoid double-counting of documents in journals with both printed and electronic ISSNs, the duplicates had to be eliminated.

The next step was to search for these "predatory" ISSNs in the Scopus citation database. Once again, this search was performed using an automatic script programmed in Python. The search was performed in March 2018. For each ISSN detected in Scopus, the script downloaded basic data on the number of documents in the "article" categories for the years 2015-2017, as well as more detailed data on the number of these articles by author's country of origin. The search request in Scopus follows, where xxxx-yyyy stands for searched ISSN:

ISSN(xxxx-yyyy) AND DOCTYPE(ar) AND PUBYEAR > 2014 AND PUBYEAR < 2018

In total 324 individual journals were identified. 287 are attributed to the list of publishers and 37 to the list of standalone journals. This means that almost 10% of the journals in our database had at least one entry indexed in Scopus.

The data on the size of total research output are necessary to measure the penetration of predatory journals in individual countries. The search was done such that both the number of predatory articles in the country and the size of research output would be maximally comparable. So for all countries, the total number of articles in 2015 – 2017 was downloaded. The list of exact names of countries to download data for was taken from previous search for predatory journals, completed with a few countries for which we did not found any predatory article in. The search was performed using following request:

AFFILCOUNTRY(country) AND DOCTYPE(ar) AND PUBYEAR > 2014 AND PUBYEAR < 2018

The article, either predatory or not, is linked to the country when affiliation of at least one of his authors is located in the country. The number of authors from individual country reported by Scopus does not take into account the share of author on the article. Both the article created by 5 authors from single country and the one created by 5 authors from 5 different countries is counted as 1 article for each participating country.

The final dataset consist of the cross-section of almost all countries in the world. For each country both the total number of articles reported to Scopus, the number of articles belonging to Beall's list of standalone journals and the list of publishers. To account for the Frontiers controversy described above and to allow for investigation of the structure of their authors independently, the articles related to the Frontiers publisher were excluded from the list of publishers and treated as independent.

Results

For the sake of simplicity we refer to '*predatory articles'* as to all articles published in Beall's listed journals, i.e. the sum of articles from standalone journals, the journals from list of publishers and from Frontiers.

The analysis considers evidence over the period 2015-2017. As already noted above, using older data bears the risk that some of the journals currently featuring on Beall's lists were not yet predatory at that time. But we use data from the last three years, rather than only the most recent year, to increase the number of articles available and the robustness of the results.

Out of more than two hundred countries the data are available for, the dependent territories and countries with fewer than 300 thousand inhabitants were excluded. Among others, this concerns mainly a number of small island states in the Caribbean and Pacific, which are not the main focus of our analysis. As a result, the sample consists of 174 states, including a large number of developing countries, which altogether cover the overwhelming majority of the world's population and research activity.

In total 164 073 predatory articles were identified in Scopus in 2015-2017, i.e. 2.8 % of all articles added to Scopus in the respective period. 110 971 come from the list of publishers. Additional 30 867 is related to the publisher Frontiers. The list of standalone journals contains 22 235 articles indexed in Scopus. Hence the list of publishers is the dominant category, while the standalone journals are minor. For some articles Scopus report the country of origin as *Undefined*. These are neglected in the analysis. However, overwhelming majority of the articles found were still included in the cross-section analysis as only 1 069 predatory articles is reported as *Undefined* and more than 100 predatory articles are related to the countries excluded from the dataset. More than 99 % of predatory articles is included to the final analysis.

Table 1: The average tendencies to publish in predatory journals (% of all articles)

Country groups*	Average share of predatory articles on all articles in 2015-17				
	Standalone	*Frontiers*	*Other publishers*	*Total*	*Total excl. Frontiers*
All countries in dataset	0,46%	0,51%	1,59%	2,56%	2,05%
*Income groups**					
OECD countries	0,15%	0,57%	0,67%	1,39%	0,82%
non-OECD countries					
High income	0,38%	0,43%	2,05%	2,86%	2,43%
Upper middle income	0,57%	0,55%	2,04%	3,16%	2,61%
Lower middle income	0,81%	0,45%	2,15%	3,40%	2,96%
Low income	0,16%	0,54%	0,82%	1,51%	0,97%
*Regions**					
Middle East & North Africa	1,16%	0,49%	3,67%	5,32%	4,84%
Central Asia	1,00%	0,47%	2,86%	4,33%	3,86%
East Asia	0,62%	0,39%	3,01%	4,02%	3,63%
South Asia	0,62%	0,27%	1,81%	2,69%	2,42%
Sub-Saharan Africa	0,32%	0,57%	1,07%	1,96%	1,39%
Europe	0,34%	0,61%	0,98%	1,93%	1,32%
North America	0,03%	0,56%	0,55%	1,13%	0,57%
Latin America & Caribbean	0,10%	0,48%	0,55%	1,13%	0,65%
Pacific	0,04%	0,47%	0,38%	0,89%	0,42%

** Within-group cross-border co-authorships are counted multiple times both in Scopus Articles and in Predatory Articles. The resulting share thus should not be systematically biased*

Table 1 presents the average tendency to publish in predatory journals for different country groups. The country figures are calculated as a simple share of articles linked to respective Beall's list on all articles reported to Scopus in the individual country. Countries are grouped by the income and regional categories. The income grouping is derived from the most recent World Bank (2017) classification. Only OECD countries are excluded as an independent group. The same source is also used for regional groups. Only *Europe & Central Asia* and *East Asia & Pacific* are both divided on two independent regions. Because the paper focuses on cross-country heterogeneity, the simple averages are used to calculate country groups. For example Sri Lanka is weighted equally as India for within-group averages.

Generally, the predatory articles are most often published by authors from middle-income and high-income countries outside the OECD. These include developing countries on the convergence path to technological frontier such as India, Indonesia, Nigeria or Egypt. Some of these countries have more than 10 % of predatory articles. Frequent contributors to predatory journals often originate also from the high income countries that are not members of OECD. The average for this group is driven up by oil-rich countries like Oman, Brunei, Saudi Arabia, Bahrain, Kuwait and the United Arab Emirates.

Surprisingly, the opposite end of the spectrum with the lowest shares of predatory articles is also dominated by developing countries. Chad, North Korea, Afghanistan or Haiti has less than 0.4 % of predatory articles. In a few cases there are no authors publishing in predatory journals whatsoever. In other low-income countries such as Niger, Ethiopia or Zimbabwe, the tendency

to publish in predatory journals is above world average. However these are still well below the most severely hit countries and the size of research output in these countries tends to be rather small. In general, the predatory journals do not belong among major publications platforms in low income countries.

Developed countries are generally less prone to predatory publishing. With exception of South Korea and Slovakia all OECD countries have below average tendency to publish in predatory journals. For example in Netherlands, Switzerland and Denmark the tendency is below 1.5 %. Moreover approximately two thirds of predatory articles reported from these countries are published by publisher Frontiers. As expected these countries still are not perfectly immune to the predatory phenomenon.

From the regional perspective the countries severely hit are located in the Middle East & North Africa, South Asia and Central Asia, where on average more than 3.5 % of predatory articles are published in predatory journals. Many countries on the convergence path such as India, but also oil-rich countries are located in these areas. Sub-Saharan African countries, with few exception such as Nigeria or Sudan, do not belong among the most frequent contributors to predatory journals. Similarly in Latin America and Pacific countries the propensity to publish predatory articles tends to be low. The relatively low penetration of predatory articles for example in Latin America, but also in China or Francophone countries in Sub-Saharan Africa can probably be partly explained by the English bias of Beall's list.

The above described distribution of predatory publishing holds for the both the list of standalone journals and the list of publishers. As they account for majority of all predatory articles, the similar distribution is also apparent when analysing all predatory articles. However, journals of the Frontiers publisher exhibit different pattern. The tendency to publish is roughly constant around 0.5 % across income groups. When neglecting for countries with small research sector (less than 1 000 articles in Scopus in total), the highest propensities to publish in Frontiers journals are in Luxembourg, Chile, Malta, Austria, Switzerland, Netherlands, Germany, Italy, Israel or Belgium. In all these countries more than 1 % of the country's articles were published in Frontiers journals. In all these countries with exception of Italy this is much more than all other predatory journals together. At least from the geographical perspective, the Frontiers journals do not behave as a typical predatory outlet.

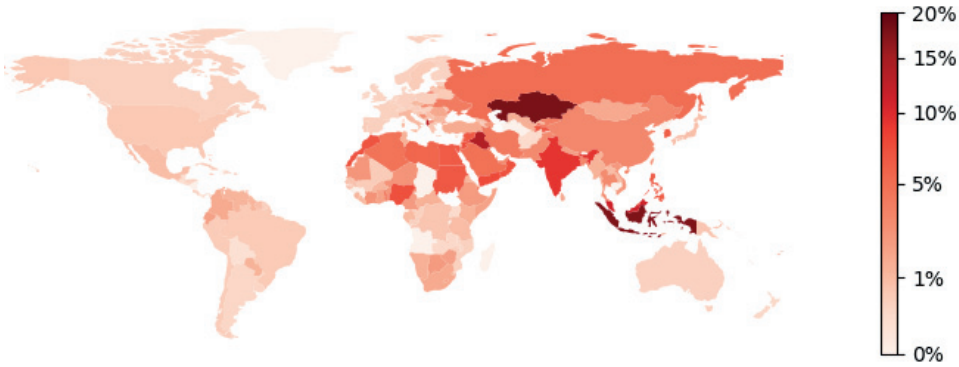


Figure 1: Share of predatory articles on total number of articles in 2015-2017
Source: Scopus, own calculation; the articles from the Frontiers publisher were excluded.

The tendency to publish in predatory journals across countries is also visualized on the world map in the figure 1. The Frontiers articles were neglected for creation of this map as they are clearly atypical. Only results related to the list of publishers and the list of standalone journals are included. The darker the color, the higher the tendency to publish in predatory journals in the country. Note the scale of the map - almost every fifth article is marked as predatory in the darkest coloured countries.

The map confirms that the most severely hit countries are concentrated in Asia and in Northern Africa. 18 out of 21 countries with more than 5 % of predatory articles (without Frontiers) are located in these regions. Also the other two rather confirm the trend – both Russia and Nigeria are not very far from the '*predatory heartland*'. In Kazakhstan, Indonesia, Iraq and Albania out of all articles reported to Scopus more than 10 % of relates to journals on Beall's list (excluding Frontiers).

A very different picture appears when looking at the predatory articles in absolute terms and ignoring the size of the research sector. Although predatory articles consist only 3.6 % of all Chinese articles, the highest number of articles originate there – 42 434. Second ranked India reported additional 28 557 predatory articles. Together with USA (16 677 articles) and South Korea (12 486 articles) these 4 countries account for half of all identified predatory articles. The authors from developed countries are still very important costumers of predatory journals. More than two fifths of all predatory articles has at least one author affiliated in OECD country. The figure decrease to approximately quarter when excluding Frontiers articles. Still every fourth predatory article originate from one of the OECD countries. Many of them from countries such United States, Germany, United Kingdom or South Korea that are considered to be at the top of the world research ladder.

Conclusions

The paper mapped the penetration of scholarly journals suspicious of using predatory practices in Beall's lists into the citation database Scopus. Using the journal's ISSNs derived from the Ulrichsweb register, Scopus search yielded 324 journals with at least one indexed article. Over the period 2015-2017, we identified 164 thousand documents in Scopus that were published in these journals, making up a share of 2.8 % of the articles in respective period. Although overall this does not look like a particularly high number, what is problematic is that in a number of countries, especially in middle income countries, the share of articles published in predatory journals turns out to fairly above the world average.

As for the cross-country differences, the results are broadly in line with the previous evidence (Bjork and Shen,2015 and Xia et al. 2015) in the sense that India comes out as one of the most affected country and that in particular Asia and Africa are the hotbeds of predatory publishing. However, the results presented in this paper provide much higher level of granularity. Several other countries outside of these regions have been shown to suffer from this problem greatly as well, such as Albania, Russia and Ukraine. Predatory publishing is a truly global phenomenon plaguing research systems across continents. It concerns any developing country that invests in research infrastructure and attempts to develop the university sector, which is the cornerstone of technological catching-up.

Without further, more detailed analysis, we can only speculate as to what exactly explains the differences we have observed in the propensity to predatory publishing across countries. However, it should not come as a surprise if this is strongly associated with the way in which research is evaluated, both on a national level in terms of the evaluation of research institutions

and project proposals, and on an individual level when deciding on career progression, and consequently indeed the way in which public money for research is allocated. It is likely that the more primitive a system a given country has for evaluating research, and the more it relies only on counting publications, the better the conditions it provides for predatory journals. To pin down the influence of these factors in a more systematic way is a major challenge for further research on this topic.

References

Bagues, M, Sylos-Labini, M and Zinovyeva, N. (2017) 'A Walk on the Wild Side: `Predatory' Journals and Information Asymmetries in Scientific Evaluations'. London, Centre for Economic Policy Research. https://www.cepr.org/active/publications/discussion_papers/dp.php?dpno=12309

Beall J. (2015) Criteria for Determining Predatory Open-Access Publishers. Scholarly Open-Access Blog.

Beall J. (2016) Scholarly Open Access: Critical analysis of scholarly open-access publishing (Beall's blog). 2016; Available from: https://scholarlyoa.com; [cited 1 April 2016].

Berger M, Cirasella J. (2015) Beyond Beall's List Better understanding predatory publishers. College & Research Libraries News. 76: 132-135.

Bloudoff-Indelicato M. (2015) Backlash after Frontiers journals added to list of questionable publishers. Nature, 526: 613.

Crawford W. (2014a) Journals, „Journals" and Wannebes: Investigating The List. Cites & Insights.; 14(7): 1–24.

Crawford W. (2014b) Ethics and Access 1: The Sad Case of Jeffrey Beall. Cites & Insights. 14(4): 1–14.

Cyranoski (2018) China awaits controversial black list of 'poor quality' journals. Nature 562, 471-472. doi: 10.1038/d41586-018-07025-5

Eriksson S., Helgesson G. (2017) The false academy: predatory publishing in science and bioethics. Med Health Care Philos 20(2), 163-170. doi: 10.1007/s11019-016-9740-3

Good B, Vermeulen N, Tiefenthaler B, Arnold E. (2015) Counting quality? The Czech performance-based research funding system. Research Evaluation 24: 91-105.

Mongeon P, Paul-Hus A. (2016) The journal coverage of Web of Science and Scopus: a comparative analysis. Scientometrics. 106(1): 213-228.

Packer AL, Cop N, Luccisano A, Ramalho A, Spinak E. SciELO – 15 Years of Open Access: an analytic study of Open Access and scholarly communication. 2014; Paris: UNESCO; Available from: http://dx.doi.org/10.7476/9789230012373; [cited 27 March 2016].

Scopus (2016). Content Policy and Selection. Available from: https://www.elsevier.com/solutions/scopus/content/content-policy-and-selection; [cited 10 October 2016].

Scimago (2016) Scimago Country Rank; Available from: http://www.scimagojr.com/countryrank.php; [cited 10 October 2016].

Shamseer et al. (2017) Potential predatory and legitimate biomedical journals: can you tell the difference? A cross-sectional comparison. BMC Medicine 15(28). doi:10.1186/s12916-017-0785-9

Shen C, Björk BC. (2015) 'Predatory' open access: a longitudinal study of article volumes and market characteristics. BMC Medicine 13(230): 1-15.

Straumsheim C. (2017) No More 'Beall's List'. Inside Higher Ed. Available from: https://www.insidehighered.com/news/2017/01/18/librarians-list-predatory-journals-reportedly-removed-due-threats-and-politics; [cited 27 March 2017].

Ulrichsweb. Ulrichsweb – Global Serials Directory. 2016; Available from: http://ulrichsweb.serialssolutions.com/; [cited 1 April 2016].

World Bank (2017). How does the World Bank classify countries; Available from: https://datahelpdesk.worldbank.org/knowledgebase/articles/378834-how-does-the-world-bank-classify-countries; [cited 10 October 2016].

Xia J, Harmon JL, Connolly KG, Donnelly RM, Anderson MR, Howard HA. Who Publishes in "Predatory" Journals? Journal of the Association for Information Science and Technology. 2015; 66(7): 1406–1417.

An Exploration on the Flow of Leading Research Talents in China: from the Perspective of Distinguished Young Scholars

Ma Tingcan[1,2], Li Ruinan[1,2], Ou Guiyan[1,2], Wu Xia[3*] and Yue Mingliang[1,2*]

[1] *matc, lirn, ougy, yueml@whlib.ac.cn*
[1] Wuhan Documentation and Information Center, Chinese Academy of Sciences, Wuhan (China)
[2] Department of Library, Information and Archives Management, University of Chinese Academy of Sciences

[3] *summerwu@whu.edu.cn*
School of Medicine, Wuhan University, Wuhan (China)

Abstract

Understanding the flow characteristics of talents is a task of great significance for talents cultivating, talent structure optimizing, talent policy making, etc. In this paper, we investigate the flow of leading research talents in China based on the academic career data of Distinguished Young Scholars sponsored by the National Natural Science Foundation of China. We explore the flow characteristics from perspectives of time, region, and institution type. We find (1) over 40% of the Distinguished Young Scholars have flow experiences and (*a*) the rate of flowing reach the highest in 5-10 years of the scholars' academic career or in 1-5 years previous to receiving the Distinguished Young Scholars Fund, (*b*) scholars are more inclined to get part-time work (as, e.g., visiting professors) as the working year grows; (2) the flow of Distinguished Young Scholars mainly occurred in the provinces with most science and education resources, with the inflow and outflow provinces becoming more diversified in recently years; (3) the flow of talents within colleges and universities is more frequent than research institutes, while research institutes are facing a severe problem of brain drain. We started up questionnaires to the Distinguished Young Scholars in hope of finding potential causes for their flows. We find that (4) the most influential factors (to talent flow) are personal research ability improvement, career development and work environment, followed by the (inflow) institution's reputation, compensation and benefits, and the (inflow) province's S&T policy environment.

Introduction

It is widely acknowledged that talent flow can contribute to the creation, dissemination and diffusion of knowledge (Czaika and Orazbayev 2018, Sugimoto et al. 2017). Understanding the flow characteristics of talents is of great significance for talents cultivating, talent structure optimizing, talent policy making, etc. (Alexander 2018, Cañibano et al. 2015). As one of the countries with largest amounts of talents, China's talent flow has attracted a lot of research interests in recent years. Based on bibliometric methods, certain researches focus on analysing the pattern of the talent flow with respect to various kinds of talents, e.g., graduate students (Borjas et al. 2018), research talents (Zhou et al. 2018), technical talents (Dai et al. 2018), innovation talents (Wei et al. 2017). Others discuss China's talent flow in the international environment (Zha 2016, Huang et al. 2017) and its effects (Lu and Zhang 2015). In Zhou et al. 2018, the authors claim that a full exploration on the flow of leading research talents in China is a very important task to carry on.

In this paper, due to the importance of the work, and also due to a real demand of certain policy makers, we investigate the flow of leading research talents in China based on the academic career data of Distinguished Young Scholars (DY Scholars) sponsored by the National Natural Science Foundation of China, to help the policy makers and the public understand the flow characteristics. Unlike the existing work that extracts geographic information from bibliometric data (that may suffer severe data cleaning problems like name disambiguation), we manually collect and clean the academic career data of Distinguished Young Scholars and carried out statistics on the data. We then explore the flow characteristics

* Corresponding Author, Email: summerwu@whu.edu.cn, yueml@whlib.ac.cn

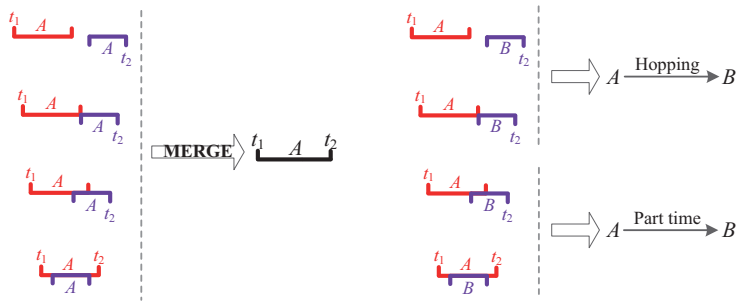
from perspectives of time, region, and institution type and finally start up questionnaires to the DY Scholars in hope of finding potential causes for their flows.

Data

The analysis is carried out based on the academic career data of DY Scholars sponsored by the National Natural Science Foundation of China (NSFC) during 1997-2011. The authors first acquire the 2422 DY Scholars' basic information (e.g., Name, Institute, Grant Year, etc.) from NSFC's official website (ISIS, https://isisn.nsfc.gov.cn/) and then manually collect their academic career data via the Internet. This manual labor is essential since there is no official website that collects the information of all the DY Scholars and each DY Scholar's homepage may have very different website styles (so that information cannot be collected using web crawler). After the comprehensive data collecting and verification, more than 6,500 education experiences and 20,000 work experiences of the 2422 DY Scholars have been acquired (last update time: Oct. 2018). The information is organized as entries with structures as follows. Basic information: [ID, Name, Institution, Grant Year, Department, Province, City, Sex, Birthplace, Birthday, Title, Position]; Education experience: [ID, University/Institute, Graduate Year, Country, Province, City]; Work experience: [ID, University/Institute, Country, Province, City, Starting Year, Ending Year, Title, Position]. That is, a total of nearly 200,000 data fields are collected and verified. After data verification, 65 DY Scholars are removed from the dataset due to incomplete information (i.e., no work experience can be found). The remaining 2357 DY Scholars are used for the flow characteristic investigation.

Method

We aim to explore the flow characteristics of DY Scholars from perspectives of time, region, and institution type. We consider 2 types of flows, i.e., job-hopping and part time job. We design a computer algorithm to determine flow types. For every DY Scholar, the algorithm first sort his (her) work experiences according to the starting and ending years in an ascending order. Thus, for each pair of consecutive work experiences, the two time intervals (corresponding to the two work experiences) at most can have four relationships, i.e., *disjoint*, *touch* (or consecutive), *overlap*, *contain* (a special case of overlap), as shown in Fig. 1. Then, for each pair of consecutive work experiences, 1) if the two institutions are the same, the algorithm merges the time intervals as a new work experience no matter what relationship occurs (Fig. 1 (*a*)), and then resorts the work experiences; 2) if the institutions are different, the algorithm concludes it as a job-hopping if the two time intervals are *consecutive* or *touch*, or as a part time job if the time intervals are *overlap* or *contain* (Fig. 1 (*b*)). The process repeats until every pair of experiences is considered.



(*a*) same instituion (*b*) different instituions

Figure 1 Processing of time intervals and determining flow types

In general, two reasons may cause the time gap between two consecutive work experiences: 1) the gap interval is not included in the DY Scholar's academic career data; 2) the DY Scholar went aboard during that interval, and we removed the experience from the dataset during data verification since in this paper we only focus on the flows happened in China. The same institution with 2 *touch*, *overlap* or *contain* time intervals is mainly caused by different titles or positions of the DY Scholar. For example, a professor working at an institution during 2001-2009 may has another work experience indicating that he was the director of the institution from 2007 to 2009.

In Fig 1(*a*), it may not that intuitive to merge an institution's two disjoint time intervals. We think it's reasonable since, 1) no information about other institutions is obtained that makes the two intervals unmergeable; 2) more importantly, under the perspective of flowing, merging the intervals will not change the flow type determination.

After the flow type determination, 1436 times of job-hopping and 571 part time jobs are found. The distribution of flows is shown in Fig. 2, in the form of number of flows, number of DY scholars (of that flow number) and the percentage with respect to the total 2357 DY Scholars. It can be seen that more than half of the DY Scholars have experienced mobility.

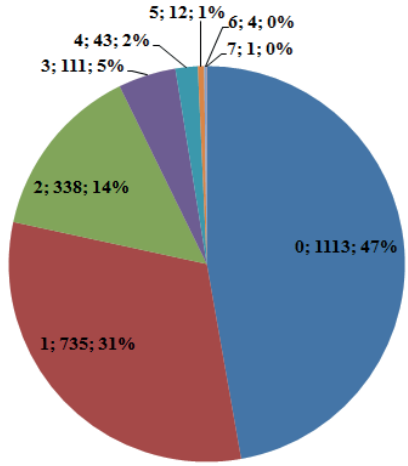


Figure 2 Distribution of numbers of flows

In the next section, we explore the flow characteristics based on the aforementioned flows from perspectives of time, region, and institution type. We make statistics on the numbers of flows by restricting flow time, outflow/inflow regions and institution types, and demonstrate and analyze the results using figures.

Flow characteristics

Time

In this section, we investigate the flow pattern from the perspective of time. For each flowing, we calculate two kinds of time intervals: 1) the time interval with respective to the first job (i.e., working time); 2) the time interval with respective to the receiving of Distinguished Young Scholars Fund.

Fig. 3 shows the number of DY Scholars that have job-hopping and part time job in each time interval based on working time, as well as the ratio of (DY Scholars having) job-hopping and part time job according to the total number of DY Scholars in each time interval (Fig. 4). In

each time interval, for every DY Scholar, multiple times of job-hopping (or part time job) are only counted once. From Fig. 4 it can be seen that almost all the DY Scholars have worked for 15 years. When considering 26 and more working years, the proportion of DY Scholars drops to 52% and less. Hence, in Fig. 3, the last two ratios corresponding to 26 years and more are plotted as dotted lines, since the statistics may change when more data can be acquired.

From Fig. 3 we can see that the number of DY Scholars having job-hopping first increases and then falls, and the same trend also appears for job-hopping ratio. The hopping number and the hopping ratio both reach the peak during the 6th-10th years of work, and then encounter an obvious decline. For part time job, however, although the number of DY Scholars that have part time job is also subject to the trend of increase-and-fall, its amplitude is much smaller, and the ratio of part time job is stable compared with job-hopping ratio. In conjunction with Fig. 5, it can be seen that as the working time increasing, the ratio of (the number of DY Scholars having) part time job to job hopping increases obviously, indicating that the DY Scholars have a preference of a *soft* flowing (still working in the original institution) than actual hopping in their later stage of academic career.

Fig. 6 gives the DY Scholars' flow distribution over time with respective to DY Scholars Fund. The overall trend of job-hopping is similar with those calculated based on working time. However, we can see much sharper fluctuations near the DY Scholars Fund. The job-hopping (ratio) bursts to a peak in 1-5 years before receiving the Fund, and encounters a huge decline in the following five years. As to part time job, a raise-and-fall can be seen near the 0-5 years after the Fund, while the other intervals are respectively stable. The statistics indicate that flowing can very likely contribute to the career development. Once certain professional accomplishments are achieved (DY Scholars in our case), which always means quite a long working time has passed (the average working time receiving DY Scholars Fund is 11 years), the willing of flowing begins to decrease.

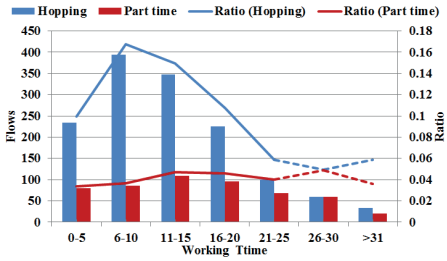


Figure 3 Flow distribution over time (working time)

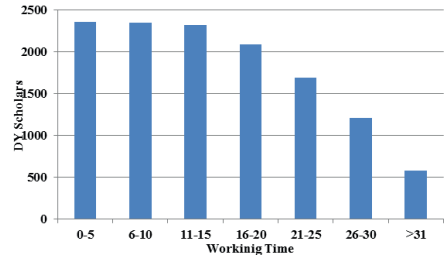


Figure 4 Distribution of total number of DY Scholars (working time)

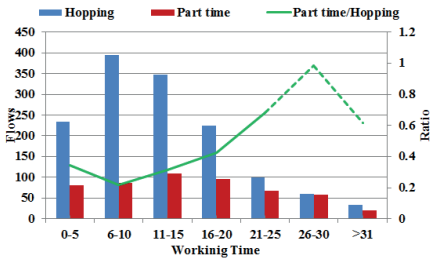


Figure 5 Ratio of part time job to job hopping (working time)

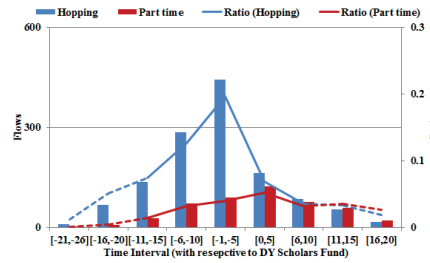


Figure 6 Flow distribution over time (with respective to DY Scholars Fund)

Region

Here we aim to investigate the flow pattern among China's 34 provinces (including municipalities). During the statistics, we count the number of inbound and outbound flows related to each province, and calculate an edge score for each pair of provinces. For each pair of provinces, by considering flows between them, we add up 1 to the edge score for each job-hopping, and 0.5 for each part time job. The edge score can be seen as a weighted summation of different flows (with different flow types) indicating the flow strength between pairs of provinces. Please note that 0.5 here is a parameter to indicate that part time job is a *soft* flowing compared to job-hopping. That means for each DY scholar, his (her) contribution to edge score is not necessarily bounded in 1. A more active scholar will contribute more to the edge score[†].

The results are shown in Fig. 7, with nodes representing provinces and edges representing flows. Provinces related to more talent flows have larger areas (Beijing has been scaled down for a better display, which is applied to all the figures in this section). Pairs of provinces that have more talent exchanges are linked with thicker lines. It can be seen that the flows of DY Scholars mainly occur in Beijing, Shanghai, Xianggang (i.e., Hong Kong), Guangdong, Jiangsu, Hubei, etc., -- the provinces that retain the most science and education resources of China.

Fig. 8 further shows the talent flows among the provinces with considering flow directions. Each directed edge in the figure represents a net inflow from the starting node to the ending node, which is calculated using the edge score of inflow minus outflow from the starting node to the ending node. The green nodes in the figure are provinces with talent net inflow, and the red nodes are those with net outflow. Larger node indicates larger net inflow/outflow. From Fig. 8 we can see that, the most popular destinations of talent flowing are Beijing, Shanghai, Guangdong, Xianggang (Hong Kong). The biggest talent output provinces are Jilin, Liaoning, Hubei, Shaanxi, Anhui, and Hunan. The fact indicates the long-term siphon effect of first-tier provinces to talents in China.

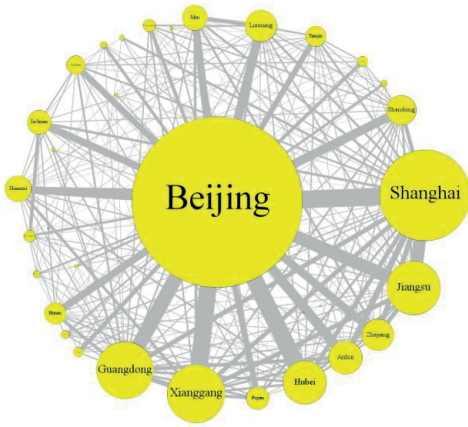


Figure 7 Flows among provinces (undirected)

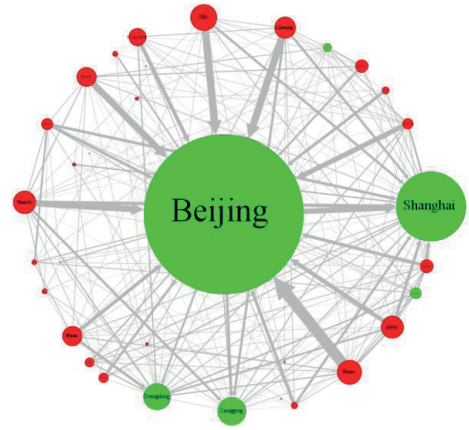


Figure 8 Flows among provinces (directed)

Fig. 9 and 10 gives the (directed) talent flow before and after 2013 respectively. By comparing the two figures we can see: 1) Beijing, Shanghai and Xianggang (Hong Kong) dominated the talent inflow before 2013, almost all the other provinces were suffering talent

[†] During the statistics, we test many values for the parameter (including 0 that means part time jobs are ignored), and the conclusion in this section is basically unchanged.

net outflow; 2) in recent five years (as of 2017), Beijing became the largest talent outflow province, and the net inflow/outflow of Shanghai and Xianggang (Hong Kong) has decreased to a very small level (hence they seem to have disappeared in the figure). On the other side, non-tier provinces (or so-called new first-tier provinces) like Hubei, Zhejiang and Shandong have become talent net-inflow provinces; 3) the number of green nodes is almost equal to the red nodes in Fig. 10, which indicates a more balanced flow of talents among provinces.

The recent changes may stem from the fact that the government has recognized the depletion of talent in non-tier provinces may further limit their development, resulting in a vicious circle. Local governments then introduced many preferential Science and Technology Policies (S&T Policies) on talent projects, entrepreneurship grants, talent-oriented housing and medical care policies and others to attract talents. The statistics in this section may show some evidence to the debate over China on whether the recent talent policies worked. And the answer is that talent flow may actually be guided by governmental policy (and the actual benefits it conveys).

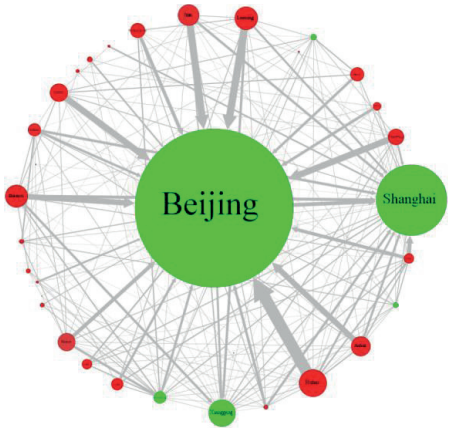


Figure 9 Flows among provinces (≤2013)

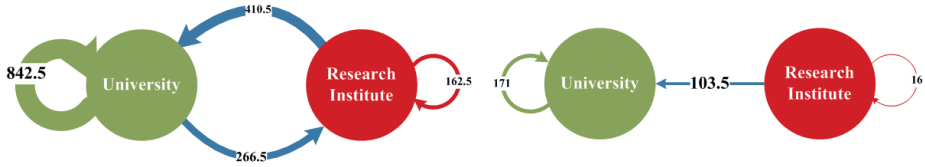
Figure 10 Flows among provinces (>2013)

Institution type

This section focuses on flows among different institution types. As the most majority (more than 95%) flows happen between university and research institute, here we only consider the two types of institutions. Fig. 11 (*a*) demonstrates the overall flows and Fig. 11 (*b*) shows the situation of the recent five years. In the figure, nodes represent different institution types, colours of nodes covey the same meaning as in the previous section (*section Region*), and areas of nodes have no meaning. Scores on the edges are weighted summations as described in the previous section.

As we can see, 1) compared with research institutes, talents in universities are more likely to flow (University 1109 vs. Institute 573); 2) the internal flow within universities is much more frequent than that within research institutes (University 842.5 vs. Institute 162.5); 3) overall research institutes have a tendency to lose talents to universities, and the situation is getting more serious -- most net outflows from research institutes to universities happened in the recent five years (i.e., after 2013, with respect to 2018).

The authors interviewed several DY Scholars in hope of grasping the reason of the situation. The (informal) answers indicate that a more flexible (research) environment, child education, and other supporting facilities are the main reasons for the flowing. Considering the different functions of research institutes and universities, the authors sincerely hope the government and the public can recognize such a trend to intervene when necessary.



(*a*) Flows (ALL) (*b*) Flows (2013-2017)

Figure 11 Flows between institution types

Questionnaires and analysis

To further understand the reasons behind the talent flows, we started up questionnaires to the DY Scholars. Questions in the questionnaires are related to Age, Sex, Birthplace, Highest Education, Research filed, School of Graduation, Age of receiving DY Scholars Fund, Years to receiving DY Scholars Fund (since work), whether flowed (in China), when flowed (years since work), factors (please refer to Tables in Appendix) that influence flowing. Questionnaires were sent to all the DY Scholars through emails. Till the statistics were carried out (January 15, 2019), 73 questionnaires were collected and used for statistics.

Certain basic statistics of the collected questionnaires are shown in Fig. 12 to 14. Fig. 12 shows the distribution of the research fields of the 73 DY Scholars. It is to be noted that the projects (including DY Scholars Projects) supported by NSFC all belong to the eight research fields shown in Fig. 12. That means the questionnaire samples (almost averagely) spread in all the research fields. Fig. 13 shows the ratio of (not) flowing, and Fig. 14 shows the flow distribution over time. It can be seen that the statistics are almost the same as Fig. 2 and 3. From Fig. 14 we can also see that averagely flow happens in the 12 years after work -- very similar with the 11 years calculated from the collected academic career data. The similarity among the statistics verifies the consistency of the academic career data and the questionnaire results, and the effectiveness of the conclusions drawn on the statistics.

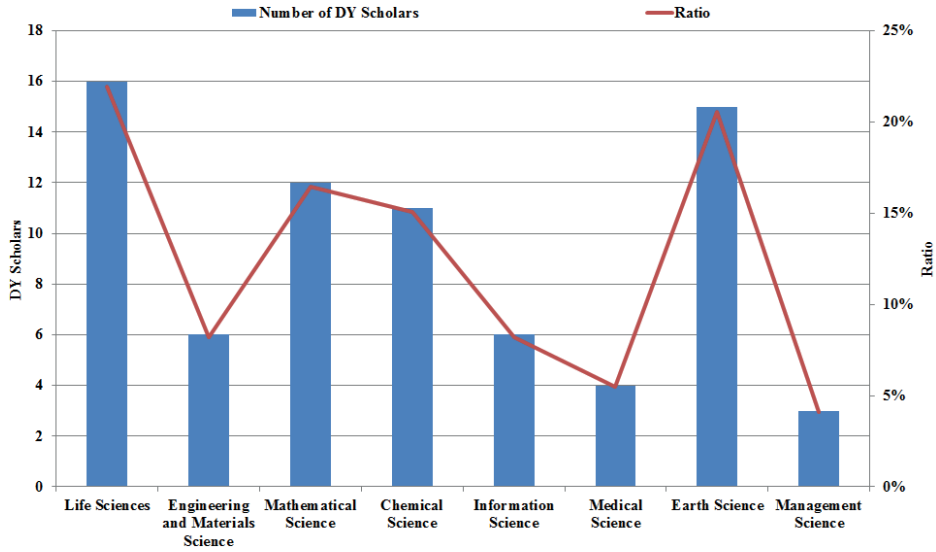


Figure 12 Research fields of the 73 DY Scholars

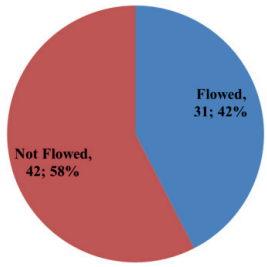


Figure 13 Flowed vs. Not Flowed

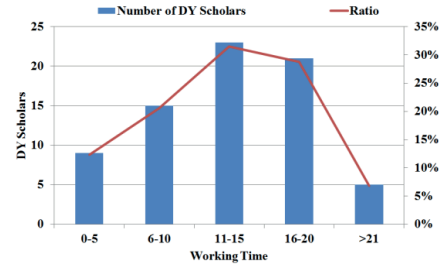


Figure 14 Flow distribution over time

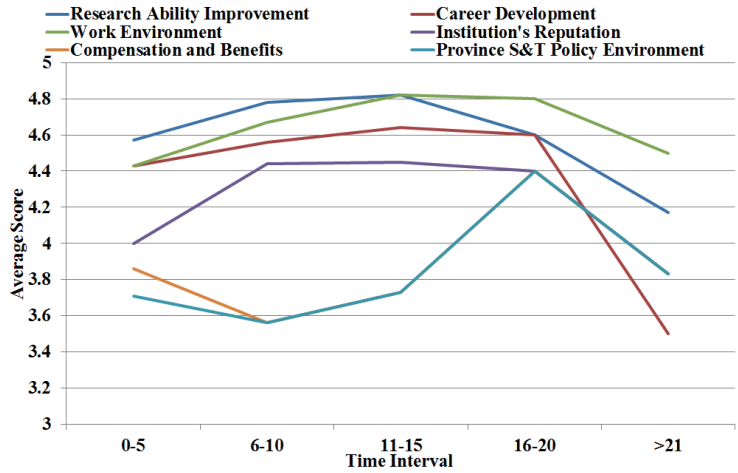


Figure 15 Total votes of each time interval

In the questionnaires, factors that influence flowing include Province Economy, Province S&T Policy Environment, Province Cultural Environment, Province Location, Province Living Environment, Compensation and Benefits, Career Development, Work Environment, Institution's Reputation, Research Ability Improvement, Family Factors and Social Relationship. For each factor, DY Scholars are asked to choose a degree to indicate how important the factor is to the flowing. The importance can be Significant, Important, General, Weak, Insignificant, each of which corresponds to a score of 5, 4, 3, 2, 1. After the voting, in each time interval, for each factor, an average score can be calculated to indicate in general, how influential the factor is to the flowing. For example, for Research Ability Improvement in Table A1, we have Average = (4*5+3*4)/(4+3)≈4.57.

Fig. 15 shows the six most influential factors, where the most influential is defined as the factors with the highest averages of the average scores over the five time intervals. From the figure we can see, first, the most influential factors are personal research ability improvement, career development and work environment. Personal research ability improvement is the most important consideration in the early ages of the academic career, with in consideration of career development. That explains why the early flowing is always accompanied by the acquisition of talent funds (see Fig. 6). In the later stage of the academic career, work environment becomes the most important factor, and the importance of career development encounters a big drop. That can somehow explain the recent talent loss from research institutes to universities: all the DY scholars get the funding before 2011, and many of them have stepped into the late stage of their academic careers -- a more flexible research

environment may be more attractive. Second, the following three factors are the (inflow) institution's reputation, compensation and benefits, and the (inflow) province's S&T policy environment. The latter two factors can be led by the government, which makes the governmental guided talent flow achievable.

Conclusion

In this paper, we collected, cleaned and analysed 2422 DY Scholars' academic career information to explore their flow characteristics from perspectives of time, region, and institution type. After statistics and analysis, the following facts were found. (1) Flowing very likely can contribute to the career development. (2) The flowing of DY Scholars is subject to the raise-and-fall trend. On the talents' later stage of academic career, part time job is (comparatively) preferable than job hopping, especially when certain professional accomplishments are achieved. (3) The flows of DY Scholars mainly occur in the provinces that retain the most science and education resources of China. (4) In recent years, non-tier provinces (or new first-tier provinces) have become talent net-inflow provinces, and the talent flow is more balanced. (5) Research institutes have a tendency to lose talents to universities and the trend is getting more serious. We also started up questionnaires to the DY Scholars in hope of finding potential causes for their flows. We found that (6) the most influential factors are personal research ability improvement, career development and work environment; (7) personal research ability improvement is the top consideration in the early ages of the academic career, while work environment becomes the most important factor in the later stage of the academic career. We hope the conclusions can help the government and the public understand the current situation and trend of talent flow, so that intervene can be taken out when necessary.

Talent flow is a very complicated social issue. The reason for a talent flow may relate to many aspects, including self-achievement, social reasons (e.g., policies), family reasons (child education), and so on. The research work in this paper was motivated by real demand of policy makers (so that only flows in China are considered). In the future work, more academic career data may be taken into account (e.g., to include foreign experience) so that more complicated patterns can be found and analyzed. Besides, there are still many DY Scholars have not replied our questionnaires. Analyzing flow characteristics with the help of more compressive survey results is also one of our future works.

Acknowledgments

This work is supported by the National Natural Science Foundation of China under grant (No. 71603252); Youth Innovation Promotion Association of Chinese Academy of Sciences (No. 2019176); Youth Innovation Team Project of National Science Library, Chinese Academy of Sciences; Teaching Research Project of School of Medicine, Wuhan University; and the Graduate Education Innovation Fund Project of National Science Library, Chinese Academy of Sciences (No. 1721-5). We'd like to show our appreciation to the anonymous DY Scholars for their kindly help on filling the questionnaires. Their answers help a lot to us on analysing the reasons behind the talent flows.

References

Alexander M. P. (2018). Multiscale impact of researcher mobility. *Journal of the* Royal Society Interface, online, DOI:https://doi.org/10.1098/rsif.2018.0580.

Borjas, G. J., Doran, K. B., & Shen, Y. (2018). Ethnic Complementarities after the Opening of China How Chinese Graduate Students Affected the Productivity of Their Advisors. Journal of Human Resources, 53(1), 1-31.

Cañibano, C., Fox, M. F., & Otamendi, F. J. (2015). Gender and patterns of temporary mobility among researchers. Science and Public Policy, 43(3), 320-331.

Czaika, M., & Orazbayev, S. (2018). The globalisation of scientific mobility, 1970–2014. Applied Geography, 96, 1-10.
Dai, Y., Kong, D., & Liu, S. (2018). Returnee Talent and Corporate Investment: Evidence from China. European Accounting Review, 27(2), 313-337.
Huang, Y., Zhao, L., & Wang, Y. (2017). The Mobility of Researchers between China and the US. In IEEE 2017 Portland International Conference on Management of Engineering and Technology, pp. 1-4.
Lu, X., & Zhang, W. (2015). The reversed brain drain: a mixed-method study of the reversed migration of Chinese overseas scientists. Science, Technology and Society, 20(3), 279-299.
Sugimoto, C. R., Robinson-García, N., Murray, D. S., Yegros-Yegros, A., Costas, R., & Larivière, V. (2017). Scientists have most impact when they're free to move. Nature News, 550(7674), 29.
Wei, F., Feng, N., & Zhang, K. H. (2017). Innovation Capability and Innovation Talents: Evidence from China Based on a Quantile Regression Approach. Sustainability, 9(7), 1218.
Zhou, Y., Guo, Y., & Liu, Y. (2018). High-level talent flow and its influence on regional unbalanced development in China. Applied Geography, 91, 89-98.
Zha, Q. (2016). What factors influence the direction of global brain circulation: the case of Chinese Canada Research Chairholders. Compare: A Journal of Comparative and International Education, 46(2), 214-234.

Appendix

Results of factors that influence flowing returned by 73 DY Scholars in various time intervals (0-5 years to more than 20 years since work) are shown in Table A1-A5.

Table A1 Results of factors that influence flowing (0-5 years)

Factors	Significant	Important	General	Weak	Insignificant	Average
Province Economy	1	3	2	1	0	3.57
Province S&T Policy Environment	0	5	2	0	0	3.71
Province Cultural Environment	0	5	2	0	0	3.71
Province Location	1	3	3	0	0	3.71
Province Living environment	1	5	1	0	0	4
Compensation and Benefits	1	4	2	0	0	3.86
Career Development	4	2	1	0	0	4.43
Work Environment	3	4	0	0	0	4.43
Institution's Reputation	1	5	1	0	0	4
Research Ability Improvement	4	3	0	0	0	4.57
Family Factors	1	2	1	2	1	3
Social Relationship	0	3	2	2	0	3.14

Table A2 Results of factors that influence flowing (6-10 years)

Factors	Significant	Important	General	Weak	Insignificant	Average
Province Economy	0	4	3	2	0	3.22
Province S&T Policy Environment	0	6	2	1	0	3.56
Province Cultural Environment	0	5	4	0	0	3.56
Province Location	0	5	4	0	0	3.56
Province Living environment	0	5	3	1	0	3.44
Compensation and Benefits	1	5	3	0	0	3.78
Career Development	5	4	0	0	0	4.56
Work Environment	6	3	0	0	0	4.67
Institution's Reputation	4	5	0	0	0	4.44
Research Ability Improvement	8	0	1	0	0	4.78
Family Factors	1	3	3	2	0	3.33
Social Relationship	0	3	5	1	0	3.22

Table A3 Results of factors that influence flowing (11-15 years)

Factors	Significant	Important	General	Weak	Insignificant	Average
Province Economy	2	2	6	1	0	3.45
Province S&T Policy Environment	2	5	3	1	0	3.73
Province Cultural Environment	2	5	3	1	0	3.73
Province Location	1	4	5	1	0	3.45
Province Living environment	0	5	5	1	0	3.36
Compensation and Benefits	6	2	2	1	0	4.18
Career Development	8	2	1	0	0	4.64
Work Environment	9	2	0	0	0	4.82
Institution's Reputation	5	6	0	0	0	4.45
Research Ability Improvement	9	2	0	0	0	4.82
Family Factors	3	5	3	0	0	4
Social Relationship	0	4	5	1	1	3.09

Table A4 Results of factors that influence flowing (15-20 years)

Factors	Significant	Important	General	Weak	Insignificant	Average
Province Economy	2	2	1	0	0	4.2
Province S&T Policy Environment	3	1	1	0	0	4.4
Province Cultural Environment	2	2	1	0	0	4.2
Province Location	2	2	1	0	0	4.2
Province Living environment	4	1	0	0	0	4.8
Compensation and Benefits	3	2	0	0	0	4.6
Career Development	4	0	1	0	0	4.6
Work Environment	4	1	0	0	0	4.8
Institution's Reputation	4	0	0	1	0	4.4
Research Ability Improvement	4	0	1	0	0	4.6
Family Factors	2	1	1	1	0	3.8
Social Relationship	0	4	0	1	0	3.6

Table A5 Results of factors that influence flowing (more than 20 years)

Factors	Significant	Important	General	Weak	Insignificant	Average
Province Economy	0	3	2	1	0	3.33
Province S&T Policy Environment	2	2	1	1	0	3.83
Province Cultural Environment	1	3	0	2	0	3.5
Province Location	1	2	1	2	0	3.33
Province Living environment	0	3	1	2	0	3.17
Compensation and Benefits	1	1	4	0	0	3.5
Career Development	1	2	2	1	0	3.5
Work Environment	3	3	0	0	0	4.5
Institution's Reputation	0	5	1	0	0	3.83
Research Ability Improvement	2	3	1	0	0	4.17
Family Factors	1	2	1	2	0	3.33
Social Relationship	2	1	1	2	0	3.5

Is Reference Publication Year Spectroscopy acceptable for Chinese Publications: Taking iMetrics Research in China as An Example

Xin Li

lucian@whu.edu.cn
School of Information Management, Wuhan University, Wuhan430072, Hubei (China)

Abstract

To identify the origin of a specific domain or topic, Max et. al. (2014) have suggested a bibliometrics method, the Reference Publication Year Spectroscopy (RPYS), which have been widely used in previous studies. However, the application of RPYS was limited within publications written in English. Thus, the aim of this study was to solve two problems: (1) Is RPYS acceptable for datasets beyond papers written in English, for example, Chinese publications? (2) Do academic domains in China share the same historical roots with the English world? To answer these two questions, we chose iMetrics for a case study. First, publications on iMetrics indexed by the China Social Sciences Citation Index (CSSCI) during 1998-2018, were collected. Then, the distribution of cited references of Chinese iMetrics research was investigated from two perspectives. Finally, to conduct RPYS analysis on this Chinese dataset, a software RootCite that accepts raw data from Web of Science and CSSCI was developed. The result shows that most of the cited references of Chinese iMetrics were publications written in Chinese, which should not be ignored, especially when analysing the origin of Chinese iMetrics. Moreover, it was found that sixteen peaks comprising of thirteen English publications and three Chinese publications are significant to the origin and evolution of iMetrics research in China. In conclusion, RPYS can be utilized not only for the dataset from Web of Science but also with the Chinese publications download from China Social Sciences Index. Studies in domestic and foreign both have the irreplaceable positive effects on the origin and development of iMetrics research in China.

Introduction

Algorithmic historiography originally proposed by Garfield E (1964) is an interesting topic in the fields of iMetrics. Since the explosive growth in the total number of academic papers, to effectively identify the valuable seminal works that are significant to the origin of a given research domain or topic, scientists in bibliometrics began to seek a computational framework that can complete the task in 20th century. It came true when Garfield and Pudovkin (2004) introduced HistCite^{TM}, a programme has been widely utilized to reveal the intellectual history of a given research field using citation links between publications chronologically. For example, the software has been used by Linnenluecke and Griffiths (2012) to uncover the origin of the domain of the corporate sustainability or by Elango (2017) tracing the significant works of the field of nanotechnology.

In recent years, a novel computation framework for algorithmic historiography-- Reference Publication Year Spectroscopy (RPYS)-- proposed by Max, Bornmann, Barth and Leydesdorff (2014), has drawn heavy attention. Since effective domain-normalization from the perspective from cited references and time-normalization by 5-year deviation from yearly citation median, RPYS can not only recognize the milestones in the development of a research field, but also can identify early works that published even earlier than the existence of the research domain. In the present literature, the studies relating to RPYS mainly focus on two aspects, one of which is the application of RPYS to identify the seminal works that significant to the origin and evolution of a given research domain, scientist or topic. Specifically, up to now, it has been successfully employed to research fields including global positioning system (GPS) (Comins J, & Hussey T, 2015), depression (Geraei E, Shakibaei F, & Mazaheri E, 2018), iMetircs (Leydesdorff L, 2014), tribology (Elango B, Bornmann L, & Kannan G, 2016) and the Darwin finches (Marx W, & Bornmann L, 2014). In addition, RPYS was also adopted to identifying research fronts and sleeping beauties by Comins and Leydesdorff (2016) and seminal works of individuals by Bornmann L et. al. (2018).

Another important research direction of RPYS is the development and optimization of the tools for RPYS analysis. The first tools that available for RPYS analysis was RPYS.exe developed by (Ledesdorff et. al., 2014), which can only compute the standard RPYS without the visualization function. Then, Thor et. al. (2016) introduced a user-friendly programme CRExplorer.exe with powerful graph making for standard RPYS. In the same year, Comins and Leydesdorff (2016) proposed Multi_RPYS and designed a web-based tool called RPYS i/o that can compute and visualize standard RPYS and Multi-RPYS, but it cannot process the data exceeding 15MB. To resolve the restrictions of previous tools, in most recently, McLevey and McIlroy-Young (2017) introduced a full-featured python package named metaknowledge that accepts large scale of data from academic databases. Moreover, a web-based tool called Patent citation spectroscopy (PCS) available for identifying seminal patents are introduced by Comins et. al. (2018). However, none of these tools accepts the raw data from Chinese academic databases, thus it is quite urgent that a tool should be developed for reconstructing an overview of reference publication year spectroscopy of research domains in China, a country with the largest amount of scientific papers.

Therefore, in this study, we introduce a python-based software (RootCite) that accepts bibliographic data from both Web of Science (WOS) and China Social Science Citation Index (CSSCI) for computing the value of standard RPYS of a given domain or topic in China or the world. This study committed to answering the following two research questions:

(1) Is RPYS acceptable for datasets beyond papers written in English, for example, Chinese publications?

(2) Do academic domains in China share the same historical roots with the English world?

Dataset and Methodology

Dataset of iMetrics research in China

The China Social Sciences Citation Index (CSSCI) is one of the most authority academic database with citation indexes in China that contains the most influential Chinese journals in humanities and social Sciences, so the dataset used in this study were collected from the online version of CSSCI operated by Institute for Chinese Social Sciences Research and Assessment, Nanjing, China, on Jan 16, 2019.

To guarantee the balance of recall and precision of our search results, we first collected all 2376 publications indexed by CSSCI whose title contains “bibliometrics” or “scientometrics” or “informetrics” or “webmetrics” to establish a core set of iMetrics research. Then we obtained a list of the most relevant and frequent domain-specific vocabularies of iMetrics research in China by analyzing the words of titles, abstracts and author-selected keywords of the core set. Thereafter, based on the domain-specific terms, we extended the former search query as follows (translated in English and the Chinese version can be seen in the Appendix information).

All Fields = (‘bibliometrics’ OR ‘webmetrics’ OR ‘informetrics’ OR ‘scientometrics’ OR ‘knowledge metrics’ OR ‘citation analysis’ OR ‘altmetrics’ OR ‘co-word analysis’ OR ‘journal evaluation’ OR ‘paper evaluation’ OR ‘scientific evaluation’ OR ‘academic impact’ OR ‘h index’) AND *publication year* = (all-time spans).

Finally, a total of 5213 related publications were identified from CSSCI, as shown in Table 1, there are totally 74,746 cited references in which references written in Chinese accounts for 63.1% (47,145), while the earliest cited reference was published in 1822 and written in English.

Table 1. A brief overview of our dataset in this study.

Item	*Value*
Number of publications	5, 213
Publication Year	1998-2018
Average Number of References	14.3
Total Number of References	74,746
Reference Publication Year	1822-2018
Number of References in Chinese	47,145
Number of References in English	27,601

Method and toolkit

Reference publication year spectroscopy was originally proposed by Max et. al (2014) for tracking the historical roots and seminal works of a specific domain, researcher or topic. Compared to traditional methods, such as simple citation counts or HistCite, RPYS take the negative effects of citations from other domains and the publication time into consideration by the utilization of visual peaks and 5-year deviations from the perspective from cited references. In previous studies relating to RPYS, tools including RPYS.exe, RPYS i/o, CitedReferencesExplorer (CRExplorer), and Metaknowledge were developed for conducting RPYS analysis on the dataset in the form of plain text downloaded from Web of Science, but, unfortunately, none of them can be applied to publications written in Chinese. For the purpose of this study, we designed and developed a tool called RootCite that can be treated as a Chinese version of RPYS.exe, which can be used for publications collected from both Web of Science and China Social Sciences Citation Index (CSSCI). AS we can see from Figure 1, RootCite contains four modules: (1) file module (creat); (2) preprocessing module (readCSSCI and readWOS); (3) rpys module (rpys and year) and (4) deduplication module (deduplication).

To investigate the evolution and historical roots of iMetrics research in China, the following four detached steps have been adopted with RootCite.

Step 1: **generating a new project**. Double click the RootCite to start it up and click the create button in the file module to create a new project, then you can find a folder called RootCiteProject including two subfolders (data_cssci and data_wos) in current directory.

Step 2: **preprocessing**. Put one or more plain texts downloaded from CSSI to the data_cssci folder and click the readCSSCI in the preprocessing module to extract all cited references.

Step 3: **computing the value of rpys and median**. Click the deduplication button to deduplicate the variants and misspelling of cited references, and click the rpys button, then rpys_cssci.csv and median_cssci.csv will be generated. Thereafter, click the year button, and the file result_cssci.csv will be generated.

Step 4: **visualization and analysis**. Using Excel to open median_cssci.csv, we can draw the reference publication year spectroscopy that can be seen from Figure 4 and identify peak RPYs; then we can find the details about the significant publications in a specific peak year with result_cssci.csv.

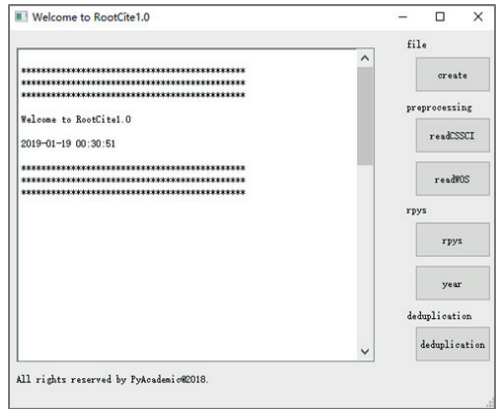


Figure 1. The main interface of the RootCite1.0

Results

The distribution of cited references from two perspectives

publication perspective

Figure 2 shows the distribution of cited references of iMetrics research in China from the publication year (1998-2017) perspective. As shown in Figure 2(A), the total number of cited references of publications related to iMetrics, dramatically increased from 474 in 1998 to more than 9300 in 2017, almost 20 times. The red curve denotes the number of cited references written in Chinese as a function of publication year, from which one can observe that, as time goes by, the number of cited references written in Chinese also increases, especially after the year 2007. One can also find that, the green line that represent the number of cited references written in English significantly grew from less than 80 in 1998 to 4366 in 2017, more than 50 times. Especially, it can be found that the red line was always above the green line overall, which illustrates that the number of the cited references written in Chinese have been more than that of cited references written in English. Moreover, the gaps between the number of cited references in two languages increased from 322 in 1998, and reached to its maximum 1771 in 2010, and then decreased to 647 in 2017.

The average number of cited references of iMetrics research in China was also analyzed according to the publication year (1998-2017). As illustrated in Figure 2(B), with the growth of the number of publications, the average number of cited references also increased steadily, from 5.1 in 1998 to 24.3 in 2017. Figure 2(B) also shows that less than 100 publications were yearly produced before 2000, which was the budding stage of the iMetrics research in China. From 2001 to 2009, which was the high-speed development stage of the Chinese iMetrics research, the annual number of publications grew rapidly. Thereafter, the produce of publications kept steadily, when the iMetrics research in China enter its maturity stage.

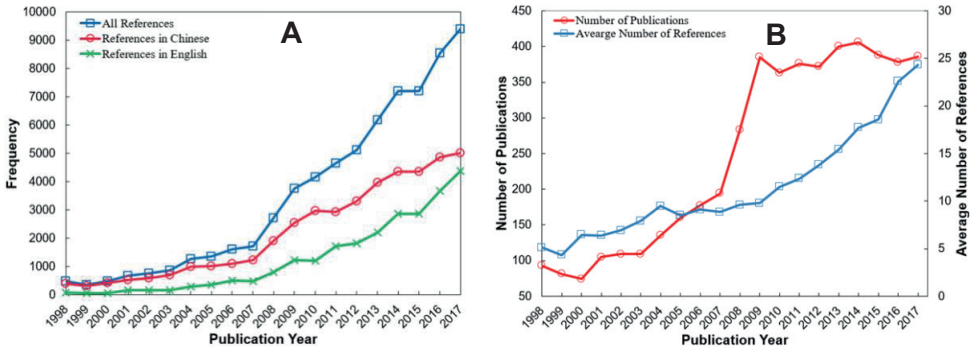


Figure 2. (A) The annual distribution of references in two languages of the iMetrics papers published in China (1998-2017); (B) The distribution for publications and references per paper of iMetrics research in China according to the publication year (1998-2017).

References publication perspective

Figure 3 shows the distribution of cited references of the iMetrics research in China with a 100% stacked area graph, in which the cited references written in two different languages respectively as a percentage of all cited references of the iMetrics research in China is presented. one can find that before 1960, the percentage of cited references in two languages fluctuated dramatically and no obviously trends is found, since the number of publications and their references was very small. From 1961 to 1982, the blue area denotes that the percentage of cited references written in English was far more than that of cited references written in Chinese represented by the red area, which indicates that, at the early stage of the Chinese iMetrics research, pioneers in China tended to absorb the advanced experience from abroad. At 1983, the percentage of cited references written in Chinese firstly exceeded 50% and kept to 2000, which reflects the development of the iMetrics research in China. Moreover, the percentages of Chinese references have been above 60% after 2000, when the Chinese iMetrics research experienced its high-speed development stage and entered its maturity.

In a word, considering the large amount and percentage of cited references written in Chinese, it can be concluded that the contribution of China should not be ignored, and it is necessary to take into account the Chinese dataset for identifying the origin and evolution of the iMetrics research in China.

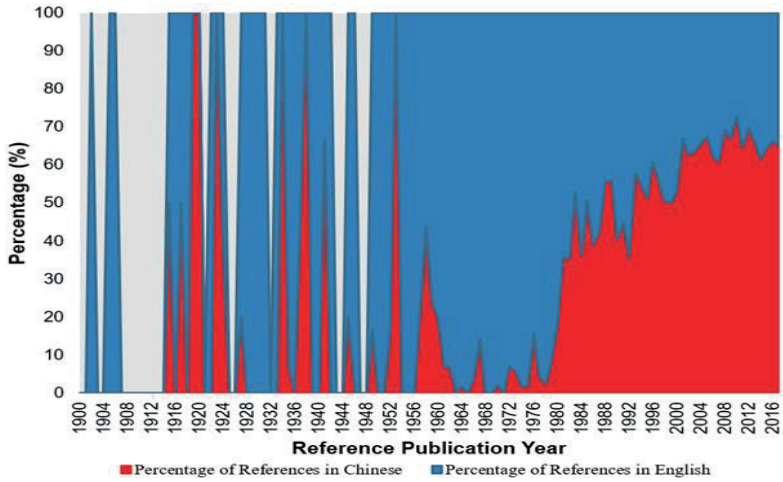


Figure 3. The percentage of references in two languages of iMetrics research in China according to the reference publication year (1900-2017).

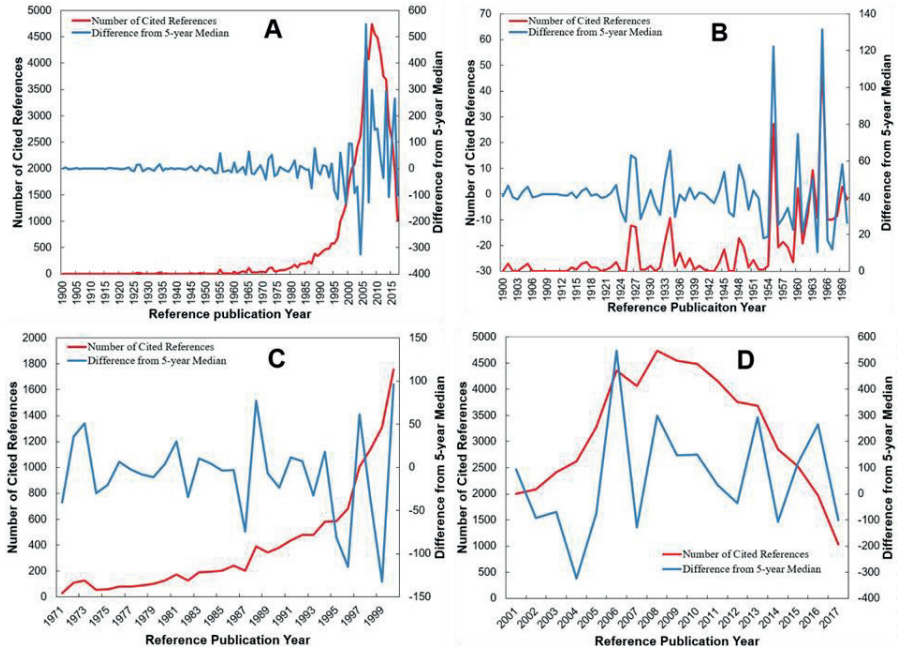


Figure 4. The Reference Publication Year Spectroscopy of iMetrics research in China organized by the reference publication year: (A) 1900-2017; (B) 1900-1970; (C) 1971-2000; (D) 2001-2017.

The reference publication year spectroscopy of iMetrics research in China

We employed RootCite to conduct RPYS analysis on the dataset of Chinese iMterics research downloaded from CSSCI. The standard RPYS graph of iMerics studies in China is presented in Figure 4 that are organized by the reference publication year: (A) 1900-2017; (B) 1900-1970; (C) 1971-2000; (D) 2001-2017. As illustrated in Figure 4(A), the RPY 2008 obtained the most cited times, which indicates the intensive relevant contribution to the Chinese iMetrics research. The blue line denotes the 5-year deviations from the median, namely that the difference from

the median of the number of the cited references in 5 years including the first two years, the current year and the next two years. One can find that the positive peaks of the blue curve densely distributed between the RPYs 1995-2008. Nevertheless, some earlier RPYs appears to be significant too, for examples, 1926, 1934 or 1988. For the purpose of describing an overview of the origin and evolution in Chinese iMetrics studies, thus we divided the RPYs into before 1970, 1971-2000 and 2000-2017.

Historical roots in twentieth century

(1) Part one: before 1970

As shown in Figure 4 (B), we can easily find that there are five major peaks (1926, 1934, 1955, 1960 and 1965) during the seven decades (1900-1970). And if we conduct a more careful analysis of the detail information in the file median_cssci.csv and result_cssci.csv, other three significant peaks can also be found, including 1917, 1944 and 1963. The details about all the eight peaks are presented in Table 2, from which we can find that all cited references were published on journals and written in English. The first peak refers to a paper published on Science Progress, in which Cole J and Eales B conducted a statistical analysis of anatomy papers (Cole, F. J., & Eales, N. B, 1917). This work accounts for the 100% citation rate and is generally considered as the first bibliometrics research in the world.

Table 2. Details about the significant peaks before 1970.

RPY	*Most Cited Reference*	*Percentage of citations (%)*	*Document Type*
1917	Cole, F. J., & Eales, N. B. (1917). The history of comparative anatomy: Part I.—A statistical analysis of the literature. Science Progress (1916-1919), 11(44), 578-596	100.0	Journal
1926	Lotka, A. J. (1926). The frequency distribution of scientific productivity. Journal of the Washington academy of sciences, 16(12), 317-323.	75.0	Journal
1934	Bradford, S. C. (1934). Sources of information on specific subjects. Engineering, 137, 85-86.	86.2	Journal
1944	Gosnell, C. F. (1944). Obsolescence of books in college libraries. college & research library.	80.0	Journal
1955	Garfield, E. (1955). Citation indexes for science. Science, 122, 108-111.	98.6	Journal
1960	Burton, R. E., & Kebler, R. W. (1960). The "half-life" of some scientific and technical literatures. American documentation, 11(1), 18-22.	41.9	Journal
1963	Kessler, M. M. (1963). Bibliographic coupling between scientific papers. American documentation, 14(1), 10-25.	48.2	Journal
1965	Price, D. J. D. S. (1965). Networks of scientific papers. Science, 510-515.	69.7	Journal

The second peak (1926) and the third peak (1934) respectively first proposed Lotka's law and Bradford's law, which are known as the two most basic laws in the field of bibliometrics. Lotka (1926), accounting for the total citation rate of 75%, firstly uncovered the relationship between authors and the number of their publications, while Bradford (1934) that

were cited 27 times, are widely utilized for identifying core journals in a scientific domain. In addition, we note here that, similar with the result of (Hou, 2017), the Zipf's law that is also known as one of the most basic laws in Bibliometrics was not recognized in this study.

The fourth peak refers to 1944 with a paper on literature obsolescence written by Gosnell, in which the phenomenon of the reduction in the value of scientific literature overtime was originally investigated (Gosnell, C. F., 1944). After six years, a measurement of literature obsolescence called "half-life" drawing on a concept from the domain nuclear physics was presented by Burton and Kebler (1960), referring to the sixth peak and accounting for 41.9% of the total citations.

The fifth peak happened in the RPY 1955 in which Garfield E who was famous for the father of scientometrics published an article entitled "Citation indexes for science" on the most influential journal Science (Garfield E, 1955). This work is widely considered as the initiation of the method of citation analysis and is the foundation of the Science Citation Index (SCI), which is an important database for the iMetrics research over the world.

The seventh peak is in the RPY 1963 which is because of the creation of bibliographic coupling and its application for measuring the static correlation between two scientific papers. The more bibliographic couples exist, the more relevant the two papers are (Kessler M, 1963). The last peak happened in the year 1965, one of the two most outstanding peaks (1955 and 1965) during 1900-1971, in which Price D published an article entitled "Networks of scientific paper" on Science. In his paper, Price D pointed out that the patterns of biographic information could be utilized for detecting the essence of the scientific research front (Price D, 1965).

(2) Part two: during 1971-2000

Figure 4 (C) present the reference publication year spectroscopy of Chinese iMetrics research in the period of 1971-2000. There results illustrate that there are six significant peaks in the domain of iMetrics in China during this period. The details about these peaks are shown in Figure 4 (C) and Table 3, from which one can find that two peaks refer to cited references written in Chinese and Qiu J authored both, while the remaining are all in English.

Table 3. Details about the significant peaks between 1971-2000.

RPY	*Most Cited Reference*	*Percentage of citations (%)*	*Document Type*
1973	Small, H. (1973). Co-citation in the scientific literature: A new measure of the relationship between two documents. Journal of the American Society for information Science, 24(4), 265-269.	56.0	Journal
1981	White, Howard D., and Belver C. Griffith. "Author cocitation: A literature measure of intellectual structure." Journal of the American Society for information Science 32.3 (1981): 163-171.	28.1	Journal
1983	Callon, M., Courtial, J. P., Turner, W. A., & Bauin, S. (1983). From translations to problematic networks: An introduction to co-word analysis. Information (International Social Science Council), 22(2), 191-235.	47.3	Journal
1988	邱均平,文献计量学,北京:科学技术文献出版社,1988. *[Qiu J. (1988). Bibliometrics. Science and Technology Literature Publishing House, Beijing China.]*	47.4	Book

1997	Almind, T. C., & Ingwersen, P. (1997). Informetric analyses on the world wide web: methodological approaches to 'webometrics'. Journal of documentation, 53(4), 404-426.	23.4	Journal
2000	邱均平,信息计量学系列论文, 2000. *[Qiu J. (2000). The Series of "Informetrics".]*	35.6	Journal

The first peak in the period of 1971-2000 is 1973, in which Small H originally put forwarded the method of co-citation analysis as a measurement of the correlation between two scientific papers. Co-citation analysis, as a milestone in iMetrics research, have been widespread and successfully utilized for detecting the research fronts and hot spots of a domain or topic in natural and social sciences.

After eight years, the second peak occurred referring to the variant of co-citation that is also co-citation analysis of scientific papers but on the author levels. White H and Belver C (1973) named it as author co-citation and applied it to measure the intellectual structure of a scientific domain or topic. The next significant peak happened in 1983 and it refer to a publication published by Callon M et.al (Callon, M., Courtial, J. P., Turner, W. A., & Bauin, S., 1983). The contribution of their work is the introduction of co-word analysis, which can be treated as another variant of co-citation analysis conducted on keyword level. However, the co-occurrence relationship between keywords were also considered except for co-citation relationship in the co-word analysis.

The fourth peak happened in 1988 is especially based on the book "Bibliometrics" (translated in English) by Science and Technology Literature Publishing House, Beijing China. The author of this book is Qiu J, who have been famous for his contribution to the development of bibliometrics in China. In his book, Qiu J systematical introduced the basic theories and methodologies of the field of bibliometrics to Chinese readers. Despite most part of this book is Chinese translation of research works from abroad, it is undeniable that the book has an irreplaceable positive effective on the origin and development of iMetrics research in China. Later the "Bibliometrics" was reprinted in 1983, 1985, 1988 and was used by more than 10 universities as teaching materials.

The next peak in 1997 dates to a paper "Informetric analyses on the world wide web: methodological approaches to 'webometrics'" by Almind T and Ingwersen P (1997) where the definition and method of webmetrics was proposed. Intrinsically, webmetrics was a variant of scientometrics or bibliometrics in the new era of Internet.

Finally, the last peak in the period is RPY 2000, referring to a series of papers in Chinese by Qiu J. In these papers, Qiu J introduced the definition, development and evolution of the field of informetrics. Later the series published in different Chinese journals was integrated into a book named "Informetrics" by Wuhan University Press in 2007.

Milestones in 21st century: period of 2001-2017

In 21st century, the iMetrics research in China began to expansionary development, the number of publications, cited references, citations all increased dramatically. The total number of cited references during the period of 2001-2007 is 54570 after deduplication. Although the citation times of references has not been steady, we note here that, there are two obvious peaks in the Figure 4 (D) have happened respectively in 2006 and 2008, which are significant to the development of the iMetrics research in China. Table 4 shows the detail information about the seminal works in the peaks during 2001-2017. Note that here we use cited times of cited references instead of the percentage, since large number of references' citations during this period are only one time which made the value of percentage very low.

Table 4. Details about the significant peaks between 2000-2017.

RPY	*Most Cited Reference*	*Cited times*	*Document Type*
2006	Chen, C. (2006). CiteSpace II: Detecting and visualizing emerging trends and transient patterns in scientific literature. Journal of the American Society for information Science and Technology, 57(3), 359-377.	196	Journal
2008	苏新宁. (2008). 构建人文社会科学学术期刊评价体系. 东岳论丛, (1), 35-42. *[Su X. N. (2008). Constructing an evaluation system for academic journals in humanities and social sciences. Dongyue Tribune, (1), 35-42.]*	111	Journal

The most cited paper in 2006 was "CiteSpace II: Detecting and visualizing emerging trends and transient patterns in scientific literature" by Chen C, which was published in Journal of the American Society for information Science and Technology and cited by 196 times, only 4.5% of the total citations of references published in 2006. CiteSpace, the most popular tool for knowledge mapping employed in China, was introduced in Chen's paper. The sencond peak happened in RPY 2008, referring to Su X's "Constructing an evaluation system for academic journals in humanities and social sciences", in which the construction of China Social Sciences Citation Index (CSSCI) was introduced. CSSCI has become one of the most authority academic database for humanities and social sciences in China and is also a significant data source for Chinese iMetrics research.

Discussion and conclusions

In this study, we introduced a new software, RootCite, for conducting standard RPYS analysis on publications from CSSCI and WOS. By taking Chinese iMetrics research as an example, we first plotted the distribution of the cited references of this field and found that references written in Chinese have been always more than English references, indicating the necessity of taking Chinese references into consideration when identifying the historical roots and evolution of Chinese research domains. This is especially true for recognizing the milestones in the development stage of a given field. Moreover, the outputs of RootCite show that there are totally sixteen peaks referring to thirteen English publications and three Chinese publications, demonstrating that RPYS is also acceptable for Chinese publications. In the meantime, we can find that the iMetrics research in China has almost the same origin peaks as in the worlds, as described in (Leydesdorff L et. al., 2014). In addition, despite most peaks happened abroad, there were still several Chinese works that have an irreplaceable and positive effects on the development and evolution of iMetrics research in China.

Acknowledgments

This work was funded by the Major Project of the National Social Science Foundation of China (17&ZDA292) and the National Natural Science Foundation of China (71473183). The author also gratefully acknowledges financial support from China Scholarship Council.

References

Almind, T. C., & Ingwersen, P. (1997). Informetric analyses on the world wide web: methodological approaches to 'webometrics'. *Journal of documentation*, 53(4), 404-426.

Bornmann, L., Haunschild, R., & Leydesdorff, L. (2018). Reference publication year spectroscopy (RPYS) of Eugene Garfield's publications. *Scientometrics*, 114(2), 439-448.

Bradford, S. C. (1934). Sources of information on specific subjects. *Engineering*, 137, 85-86.

Burton, R. E., & Kebler, R. W. (1960). The "half-life" of some scientific and technical literatures. *American documentation*, 11(1), 18-22.

Callon, M., Courtial, J. P., Turner, W. A., & Bauin, S. (1983). From translations to problematic networks: An introduction to co-word analysis. *Information (International Social Science Council),* 22(2), 191-235.

Chen, C. (2006). CiteSpace II: Detecting and visualizing emerging trends and transient patterns in scientific literature. *Journal of the American Society for information Science and Technology*, 57(3), 359-377.

Comins, J. A., & Hussey, T. W. (2015). Detecting seminal research contributions to the development and use of the global positioning system by reference publication year spectroscopy. *Scientometrics*, 104(2), 575-580.

Comins, J. A., & Leydesdorff, L. (2016). RPYS i/o: software demonstration of a web-based tool for the historiography and visualization of citation classics, sleeping beauties and research fronts. *Scientometrics*, 107(3), 1509-1517.

Comins, J. A., Carmack, S. A., & Leydesdorff, L. (2018). Patent citation spectroscopy (PCS): Online retrieval of landmark patents based on an algorithmic approach. *Journal of Informetrics*, 12(4), 1223-1231.

Cole, F. J., & Eales, N. B. (1917). The history of comparative anatomy: Part I.—A statistical analysis of the literature. *Science Progress* (1916-1919), 11(44), 578-596

Elango, B. (2017). Scientometric analysis of Nature Nanotechnology. *Library Hi Tech News*, 34(1), 23-30.

Elango, B., Bornmann, L., & Kannan, G. (2016). Detecting the historical roots of tribology research: a bibliometric analysis. *Scientometrics*, 107(1), 305-313.

Garfield, E., Sher, I.H., & Torpie, R.H. (1964). The use of citation data in writing the history of science. *Philadelphia: The Institute for Scientific Information, Report of research for Air Force Office of Scientific Research under contract* F49(638)-1256.

Garfield, E., & Pudovkin, A. I. (2004, November). The HistCite system for mapping and bibliometric analysis of the output of searches using the ISI Web of Knowledge. *In Proceedings of the 67th annual meeting of the American society for information science and technology* (Vol. 83).

Garfield, E. (1955). Citation indexes for science. *Science*, 122, 108-111.

Geraei, E., Shakibaei, F., & Mazaheri, E. (2018). Depression: Detecting the historical roots of research on depression prevention with reference publication year spectroscopy. *International Journal of Preventive Medicine*, 9(1), 53.

Gosnell, C. F. (1944). Obsolescence of books in college libraries. *college & research library*.

Hou, J. (2017). Exploration into the evolution and historical roots of citation analysis by referenced publication year spectroscopy. *Scientometrics*, 110(3), 1437-1452.

Kessler, M. M. (1963). Bibliographic coupling between scientific papers. *American documentation*, 14(1), 10-25.

Leydesdorff, L., Bornmann, L., Marx, W., & Milojević, S. (2014). Referenced Publication Years Spectroscopy applied to iMetrics: Scientometrics, Journal of Informetrics, and a relevant subset of JASIST. *Journal of Informetrics*, 8(1), 162-174.

Linnenluecke, M. K., & Griffiths, A. (2013). Firms and sustainability: Mapping the intellectual origins and structure of the corporate sustainability field. Global Environmental Change, 23(1), 382-391.

Lotka, A. J. (1926). The frequency distribution of scientific productivity. *Journal of the Washington academy of sciences*, 16(12), 317-323.

Marx, W., Bornmann, L., Barth, A., & Leydesdorff, L. (2014). Detecting the historical roots of research fields by reference publication year spectroscopy (RPYS). *Journal of the Association for Information Science and Technology*, 65(4), 751-764.

Marx, W., & Bornmann, L. (2014). Tracing the origin of a scientific legend by reference publication year spectroscopy (RPYS): the legend of the Darwin finches. *Scientometrics*, 99(3), 839-844.
McLevey, J., & McIlroy-Young, R. (2017). Introducing metaknowledge: Software for computational research in information science, network analysis, and science of science. *Journal of Informetrics*, 11(1), 176-197.
Price, D. J. D. S. (1965). Networks of scientific papers. *Science*, 510-515.
Qiu J. (1988). Bibliometrics. Science and Technology Literature Publishing House, Beijing China.
Small, H. (1973). Co-citation in the scientific literature: A new measure of the relationship between two documents. *Journal of the American Society for information Science*, 24(4), 265-269.
Su X. N. (2008). Constructing an evaluation system for academic journals in humanities and social sciences. *Dongyue Tribune*, (1), 35-42.
Thor, A., Marx, W., Leydesdorff, L., & Bornmann, L. (2016). Introducing CitedReferencesExplorer (CRExplorer): A program for reference publication year spectroscopy with cited references standardization. *Journal of Informetrics*, 10(2), 503-515.
White, Howard D., and Belver C. Griffith. "Author cocitation: A literature measure of intellectual structure." *Journal of the American Society for information Science* 32.3 (1981): 163-171.

Appendix Information

Search strategy in Chinese:

所有字段=（‘文献计量’ OR ‘网络计量’ OR ‘信息计量’ OR ‘科学计量’ OR ‘知识计量’ OR ‘引文分析’ OR ‘补充计量学’ OR ‘替代计量学’ OR ‘共词分析’ OR ‘期刊评价’OR ‘论文评价’ OR ‘科研评价’ OR ‘学术影响力’ OR ‘H 指数’）AND 出版年份=（1998-2018）。

Open Peer Review: The Current Landscape and Emerging Models

Dietmar Wolfram[1], Peiling Wang[2] and Hyoungjoo Park[3]

[1] *dwolfram@uwm.edu;* [3] *park32@uwm.edu*
School of Information Studies
University of Wisconsin-Milwaukee
Milwaukee, WI 53211, USA

[2] *peilingw@utk.edu*
School of Information Sciences
University of Tennessee, Knoxville
Knoxville, TN 37996, USA

Abstract

Open peer review (OPR) is an important innovation in the open science movement. OPR can play a significant role in advancing scientific communication by increasing its transparency. Despite the growing interest in OPR, adoption of this innovation since the turn of the century has been slow. This study provides the first comprehensive investigation of OPR adoption, its early adopters and the implementation models used. We identified 174 current OPR journals and analysed their wide-ranging implementations to derive emerging OPR models. The findings suggest that: 1) there has been a steady growth in OPR adoption since 2001 when 38 journals initially adopted OPR; 2) OPR adoption is most prevalent in medicine and the natural sciences; 3) three publishers are responsible for 87% of identified OPR journals; 4) early adopter publishers have implemented different models of OPR resulting in different levels of transparency. Across the variations in OPR implementations, two important factors define the degree of transparency: open identities and open reports. Open identities may include reviewer names and affiliation as well as credentials; open reports may include timestamped review histories consisting of referee reports and author rebuttals. When and where open reports can be accessed are also important factors indicating the OPR transparency level. Dimensions that characterize the observed OPR models are outlined.

Introduction and Literature Review

Peer review has been a critical process in scholarly communication. The mainstream peer review systems in scientific and scholarly communication, which typically operate anonymously (Kriegeskorte, 2012), have been criticized for being a flawed process (Smith, 2006) or broken system (Belluz, Plumer & Resnick, 2016). Peer review bias and unfairness exist to various degrees in different disciplines (Lee, Sugimoto, Zhang, and Cronin, 2013; Rath & Wang, 2017). The e-publishing era has also witnessed serious contemporary problems, among others, "predatory" open access (OA) journals as reported in Bohannon's experiment (2013) and a "peer review ring" scandal resulting in the retraction of 60 articles by a prestigious publisher's journal (Barbash, 2014).

As a contrast to the traditional, closed-peer review system, open peer review (OPR) pursues openness and transparency in the process of peer review by making the identities of the author and the reviewer of the manuscript known to each other and/or making available review reports alongside a paper or as separate entities linked to the paper. Transparency in peer review is not a new idea. It was rigorously studied by researchers for the journal *BMJ* in the 1990s. The researchers found that making reviewer identities known to authors or posting reviewer names with the paper had no effect on the quality of the reviews (Godlee, Gale, & Martyn, 1998; van Rooyen, Godlee, Evans, Black, & Smith, 1999). If transparency in peer review is the key to tackling the various issues facing the current peer review system, will authors and reviewers embrace OPR?

Launched in 2001, the journal *Atmospheric Chemistry and Physics*, was the among the first open access OPR journals (Pöschl & Koop, 2008), along with 36 journals published by BioMed Central (https://www.biomedcentral.com/journals-a-z). Since then, a small number of studies

have investigated author and reviewer attitudes towards OPR, characteristics of open reviews and methods of OPR adoption by existing and new journals.

In a large-scale international study of researchers' attitudes towards peer review, Mulligan, Hall, and Raphael (2013) found that only 20% of the respondents were in favour of making the identity of the reviewers known to authors of the reviewed manuscripts; 25% of the respondents were in favour of publishing signed review reports. In 2016, the OpenAIRE consortium conducted a survey of OPR perceptions and attitudes by inviting respondent participation through social media, distribution lists, and publishers' newsletters. Of the valid 3,062 responses, 76% of the respondents reported having taken part in an OPR process as an author, reviewer, or editor. The survey results show that the respondents are more willing to support open reports (59%) than open identity (31%). The majority of the respondents (74%) believe that the reviewers should be given the option to make their identities open. (Ross-Hellauer, Deppe & Schmidt, 2017) Another survey of European researchers conducted by the European Union's OpenUP Project in 2017 received 976 valid responses. The results of this survey also show that the respondents support open reports (39%) more than open identities (29%). This survey also reports a gender difference in supporting open identities (i.e., 35% female researchers versus 26% male researchers) (Görögh, Schmidt, Banelytė, Stanciauskas & Woutersen-Windhouwer, 2017).

In 2012, Elsevier began a pilot project to examine open review on selected trial journals (Mehmani & van Rossum, 2015). A survey of editors, authors and reviewers of the five participating trial journals was conducted in 2015 to assess the impact of open review (Mehmani, 2016). There were encouraging results. Forty-five percent of the reviewers revealed their identity. The majority of the reviewers (95%) commented that publishing review reports had no influence on their recommendations. Furthermore, 33% of the editors identified overall improvement in the review quality and 70% of these editors said that the open review reports were more in depth and constructive. Only a small fraction of the authors indicated that they would prefer not to publish in open review journals. Mehmani reported high usage of review reports by counting the clicks to the review reports, which indicated the value of open review to the readers. Some of the findings from Elsevier's pilot project corroborate other published studies on the characteristics of OPR comments and author/reviewer. Bornmann, Wolf, and Daniel (2012) compared the reviewer comments of a closed peer review journal and an open peer review journal. They found that the reviewer comments in the open review journal were significantly longer than the reviewer comments in the closed review journal. Wang, You, Rath, and Wolfram (2016) analysed the optional OPR journal *PeerJ*'s publicly available reports for the first three years of the journal (2013-2016). They found that the majority of the papers (74%) published during this time had peer review histories alongside the articles; of the published review reports, 43% included the reviewers' identities.

Vrana (2017) collected data from the websites of the top 100 scientific publishers to identify if the publishers have adopted and implemented OPR. Vrana found only nine OPR publishers, of which six listed 12 OPR journal titles. Wang and Tahamtan (2017) searched the Directory of Open Access Journals (https://doaj.com) and followed the literature and publishers of known OPR journals. They identified 155 OPR journals, of which the majority were in medicine and related fields. They also found the various characteristics in the implementations by the OPR journals.

At the 2016 Annual Meeting of the Association for Information Science and Technology, a panel of well-known scientists and editors engaged in a conversation and debate with conference attendees on the emerging open peer review innovation in the era of open science (Wang & Wolfram, 2016). Similarly, at the 8th Peer Review Congress (2017), leaders in academic publishing held a panel on "Transparency in Peer Review." The panellists discussed the various shades or spectrum of transparency in open peer review practices. Also touched

upon was the lack of transparency in research proposal reviews, especially for private foundations. Attendees at the Congress raised another important question: should there also be transparency to review reports of rejected manuscripts if they are a part of the scholarly ecosystem?

Despite the growing interest in OPR, there still is no uniform definition of OPR or generally agreed upon best implementation model. Ford (2013) reviewed the literature on the topic to define and characterize OPR. Acknowledging the diverse views of OPR, she defined OPR as "the process incorporates disclosure of authors' and reviewers' identities at some point during an article's review and publication" (p. 314). She further characterized OPR by *openness* (i.e., signed review, disclosed review, editor-mediated review, transparent review, and crowd-sourced/public review), and *timing* (pre-publication, synchronous, and post-publication). Fresco-Santalla and Hernandez-Perez (2014) illustrated how OPR has been manifested by different journals: open reviews (for all or specific papers), signed reviews (obligatory, pre- or post- publication), readership access to review reports (required or optional), readership commenting (pre- or post- publication). According to Tattersall (2015), there were ten leading OPR platforms.

Ross-Hellauer (2017) conducted a systematic literature review and identified seven elements based on 22 definitions. They defined two core elements of OPR focusing on *open identities* and *open reports*. The other five elements in the order of frequency of occurrences include *open participation*, *open interaction*, *open pre-review manuscripts*, *open final-version commenting*, and *open platforms/decoupled review*. These elements formed a framework for two surveys conducted by OpenAIRE (Ross-Hellauer, Deppe & Schmidt, 2017) and OpenUP (Görögh, Schmidt, Banelytė, Stanciauskas & Woutersen-Windhouwer, 2017). Similarly, Tennant et al. (2017) provided a comprehensive review of journals' peer review practices from the past to the present, which they published in the OPR journal *F1000Research*. Taking a much broader perspective, they examined pros and cons of open reviews including public commentary and staged publishing.

Another related development that provides credit for peer reviewers that may also have an impact on OPR adoption are services that encourage researchers to archive their peer review reports in scholarly repositories or networks such as Publons (https://publons.com/). Publons does an excellent job of authenticating review claims, but the majority of the verified reviews are not accessible due to required permissions by the journals.

Will OPR become a mainstream scholarly practice similar to open access and open data in open science? Further research is needed to understand the concept of OPR and its diverse implementations by publishers as well as the perceptions and attitudes of scientists as authors and reviewers. The purpose of this study is to conduct a thorough search for and analysis of current OPR journals to address the following research questions:

1. What is the current state of OPR?
2. What has been the trend for OPR adoption?
3. Who are the early adopters of OPR?
    a. Which disciplines have adopted OPR?
    b. Which publishers are the front runners or leaders in OPR adoption?
4. What are the emerging OPR model implementations? More specifically, what are the decision factors influencing open identities and open reports?

This study serves as the first stage of a two-phase investigation examining the current state and characteristics of OPR.

Method

As there is no comprehensive list of current OPR journals, relevant journals were identified using multiple search strategies. The Directory of Open Access Journals (DOAJ) indexes more

than 12,000 open access journals and identifies the peer review process of the journals it indexes. A search was conducted for journals identified as "open peer review." This list served as the core of the studied journals. A broader Internet search using the terms "open peer review" and "journal" was conducted using Google to identify additional titles. A third strategy was to review the literature for studies of OPR journals that were not included in DOAJ or the broader search, and by using a snowball searching technique on publisher websites to identify additional titles not found by the other approaches. In order to qualify for consideration, journals had to demonstrate adherence to at least one of two core OPR elements identified by Ross-Hellauer (2017): *open identities*, where reviewer names were made public and/or *open reports*, where the original reviews or summaries of the reviews were publicly available.

Journal data were initially collected during the summer of 2018 and updated up to December 2018. In defining the scope of OPR, we did not include journals that were limited to post-publication peer review, as these contributions may take the form of reader comments appearing after the article on the journal website. As a result of our initial searches, we found more than 230 journals. Several of the identified journal had discontinued publication and were removed from further consideration. Some journals (e.g., *BMJ Pediatrics Open* and several journals published by Copernicus Publications) indicated in their editorial policy that they follow OPR. However, if there was no evidence to support OPR (e.g., open reports or reviewer identities) in the published articles, these journals were also excluded. This exclusion extended to journals where reviewers were made known to manuscript authors during the review process but were not included in the final published version, thereby remaining hidden to readers. Some DOAJ entries for journals were blogs rather than venues for the publication of research and were also excluded. This study did not include journals that implemented only one of the following OPR elements defined by Ross-Hellauer (2017): open participation, open interaction, open pre-review manuscripts, open final-version commenting, and open platforms/decoupled review. The final list consisted of 174 OPR journals (see Appendix). Journals with asterisks represent the earliest adopters that began OPR adoption in 2001.

The DOAJ-listed information and the journal peer review policy on each journal's website were analysed to determine the accuracy of DOAJ-provided information and the extent of OPR use. Journal data were stored in an Excel spreadsheet and analysed using cross-tabulations and qualitative assessment of relevant journal content. Stored information included: journal metadata, year of first OPR use, publisher name, publisher country, policy for reviewer identity, policy for report availability, reviewer selection policy, OPR options for authors, OPR options for reviewers, report availability (what is available, when, where) and high-level journal discipline.

Results

Descriptive Data

The growth of OPR adoption—measured either by existing or new journals—is summarized in Figure 1 by broad discipline. The journals were classified into five broad topical areas using a modified form of the DOAJ classification scheme to determine which disciplinary areas have adopted OPR. Most journals did not report when they adopted OPR or if they have always used OPR. First OPR usage was confirmed by searching early issues of the journals to identify when OPR practices began. In many cases, OPR adoption coincided with the first journal issue.

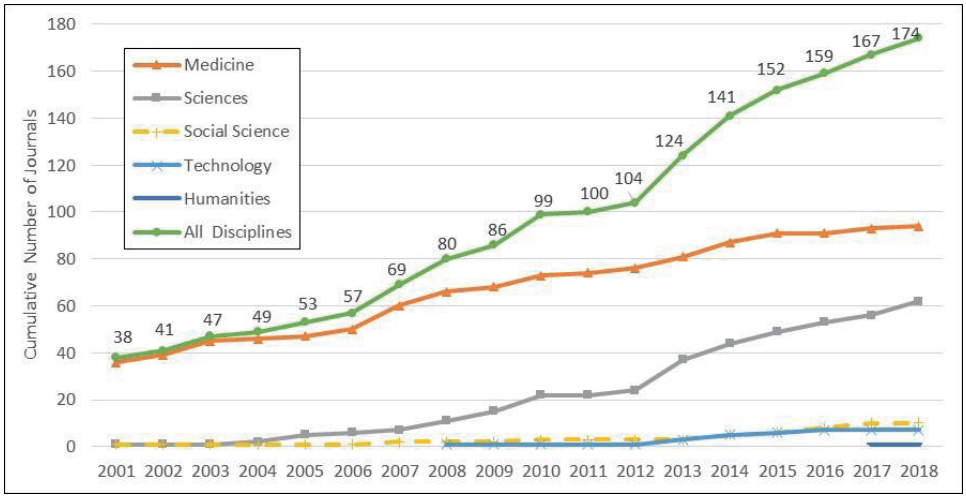


Figure 1. Growth of OPR journals by discipline groups

The early adopters of OPR can be traced back to the beginning of the 2000s. The journals *Atmospheric Chemistry and Physics* and *European Cells & Materials* each implemented a different OPR model, although both launched their first issues in 2001. Similarly, 36 OPR journals published by BioMed Central implemented another model in the same year (See the appendix for the first 38 OPR journals). Since then, there has been steady growth in the number of journals that have adopted OPR, most noticeably in medicine and, more recently, in the natural sciences over the past 10 years. The disciplinary distribution of OPR journals appears in Table 1. For each discipline group, its first OPR year and number of articles suggest how OPR as an innovation is being adopted. Medicine had the most early adopters.

Table 1. Adoption of OPR by discipline group over time

Discipline Group	*Year of First OPR Journal*	*# of OPR Journals in First Year*	*Total*	*Percentage of all OPR Journals*
Medicine	2001	36	94	54.0%
Natural Sciences	2001	1	62	35.6%
Social Sciences	2001	1	10	5.7%
Technology	2008	1	7	4.0%
Humanities	2017	1	1	0.6%
Total			174	99.9%

A summary of the most prolific OPR contributing publishers and their headquarters country appears in Table 2. Although many journals today attract an international audience and are managed by international teams of researchers, the prevalence of OPR journals associated with publishers based in Europe stands out. Of note, 87% of the OPR journals are published by three publishers (BioMed Central, Frontiers, Copernicus Publications). This points to the important role that publishers have played to date in the promotion of OPR. The 'All other publishers' category, with only one journal each, shows narrow geographic representation across 7 countries, 5 of which are also in Europe. This also points to the leading role of European publishers in this effort. All but four of the 174 OPR journals were associated with publishers based in Europe.

Table 2. Adoption of OPR by publishers

Publisher	*OPR Journals*	*OPR Articles*	*Total Journals*	*Percentage of OPR Journals*	*Headquarters Location*
BioMed Central (Springer)	68	65,771	330	20.6%	United Kingdom
Frontiers	64	95,533	64	100.0%	Switzerland
Copernicus Publications	20	39,628	38	52.6%	Germany
Elsevier	5	358	2,960	0.2%	Netherlands
F1000 Research Ltd	2	3,273	2	100%	United Kingdom
Other publishers (15)	15	7,663	--	--%	(7 countries)*
Total	174	212,226			

*Argentina (1), Bulgaria (1), Netherlands (1), Germany (1), Switzerland (1), United Kingdom (7), United States (3)

OPR in Current Practice

A fundamental principle of OPR is transparency. This includes open identities and/or open reports. Publishers and editors of journals adopted different levels of transparency, where one or both of the transparency elements may be optional or required. Table 3 reports the adoption of open reports based on the broad discipline of the journals. Approximately 63% (110/174) of the journals require or make open reports optional. The percentage is highest in medicine, and second highest in the social sciences. However, the small number of journals in social sciences means that a single journal can greatly influence the outcome. Open reports are much lower for technology and the humanities. The availability of open identities, on the other hand, was much more common. All 174 journals, except for one in the social sciences, permitted or required reviewers to identify themselves.

Table 3. Number of OPR journals adopted open reports by discipline group

Open reports / Discipline	*Available*	*OPR Journals*	*Percentage*
Medicine	69	94	73.4%
Social Sciences	6	10	60.0%
Natural Sciences	33	62	53.2%
Technology	2	7	28.6%
Humanities	0	1	0.0%
Total	110	174	63.2%

Open identities may be mandated, optional or anonymous. Similarly, open reports may be mandated, optional or not available. The frequency of each combination along with an example journal appear in Table 4. When reviewers remain anonymous and their reports are not made available, this is traditional blind peer review (the upper left cell). No examples could be found of journals that provide: 1) reviewers the option to identify themselves without making the reports available, 2) anonymous reports with optional report availability, or 3) mandated open identity with optional report availability. Examples could be found for each of the remaining categories with widely varying frequencies of implementation. The adoption of mandated open identities (141/174 or 81%) was more common than mandated open reports (107/174 or 61.5%). Fewer than half of the journals studied (77/174 or 44.3%) required that both open identities and open reports be included. Only three journals provided reviewers and authors optional open identities and optional open reports. Furthermore, more than a third of the journals (64/174 or 36.8%) published the reviewer names only with no access to the reports. Only one of the OPR journals published open reports without open identities (i.e., *Ledger*).

Table 4. Adoption of open identities and open reports

Open reports \ *Open identities*		*Anonymous*	*Optional*	*Mandated*	*Total*
None	Cases (Example)	—	—	64 (*Frontiers in Big Data*)	64
Optional	Cases (Example)	—	3 (*PeerJ*)	—	3
Mandated	Cases (Example)	1 (*Ledger*)	29 (*eLife*)	77 (*BMC Medicine*)	107
Total		1	32	141	174

Emerging OPR Implementation Models & Their Decision Factors

The current OPR landscape is complex and exhibits a variety of configurations ranging from opening some aspects of the established blind-review process to a fully transparent process. Although there is not a simple way to define the emerging OPR practices, a descriptive framework focusing on how *open identities* and *open reports* are being fulfilled (process) and what end products are available for access as depicted in Figure 2.

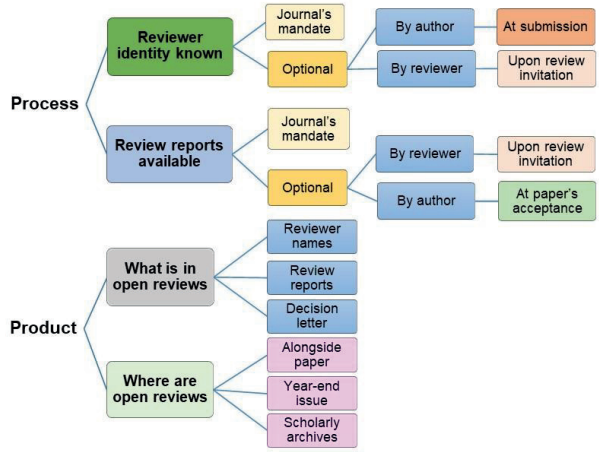


Figure 2. Process-Product Model

From a different view, various implementations models of OPR involve four factors: 1. *who* makes decisions: reviewer, author, and editor/journal; 2. *when* the decision is made for a specific core element: pre-, post, or concurrent process; 3. *what* is contained in open reports: original reports, a consolidated letter, or invited commentaries by reviewers who made significant contributions to the paper's revision; 4. *where* the open reports can be accessed.

These four factors can potentially define the level of transparency which a journal puts into practice of OPR. For example, *F1000Research* is the most transparent OPR journal because its peer review process is totally open; both reviewer identity and review comments are instantly accessible alongside of the manuscript while it is being reviewed and revised. As a contrast, the OPR journals published by Frontiers only publish each paper with its reviewers' names, which is a minimum level of open identity, although reviewers and authors interact with one another during the review process. A proposed implementation scheme, taking into consideration of the four factors, is shown in Figure 3 and illustrated below:

1. ***Who*** decides about …
   a. Open identities
   - Mandated by journal (e.g., *F1000Research*, all Frontiers journals)
   - Reviewers (e.g., *PeerJ*, *eLife*)
   - Both authors and reviewers (e.g., *Papers in Physics*)

   b. Open reports
   - Mandated by journal (e.g., all BMC OPR journals, *F1000Research*)
   - Authors of published papers (e.g., *PeerJ*)
   - Reviewers
   - Editors (e.g., *eLife*)
2. ***When*** a decision is made about …
   a. Open identities
   - At submission (e.g., all Frontiers journals)
   - Upon agreeing to review (e.g., BMC OPR journals, *F1000Research*)
   - Prior to the review process at submission and upon agreeing to review (e.g., *Papers in Physics*)
   - At submission of review report (e.g., *PeerJ*)

   b. Open reports
   - Upon manuscript being accepted for publishing (e.g., *PeerJ*)
   - Upon manuscript being accepted for publishing to selectively invite commentary from reviewers made significant contribution (e.g., *Papers in Physics*)
3. ***What*** is included in the open reviews?
   a. Original timestamped review reports (e.g., all BMC OPR journals, *F1000Research, PeerJ*)
   b. Consolidated review reports as decision letters (e.g., *eLife*)
   c. Commentary article by reviewers invited by the editor for significant contributions to the published paper (e.g., *Papers in Physics*)
   d. Names of reviewers acknowledged (e.g., Frontiers journals)
4. ***Where*** are open reports accessible?
   a. Added to the article as a section (e.g., *European Cells & Materials*)
   b. Standalone page or file alongside the publication (e.g., *PeerJ*, *eLife*, all Copernicus OPR journals)
   c. A commentary article in the same issue of the article (e.g., *Papers in Physics*)
   d. A dedicated year-end Supplement issue (e.g., Elsevier's 5 trial journals)
   e. Reviews archived in scholarly network services (e.g., *Publons*)

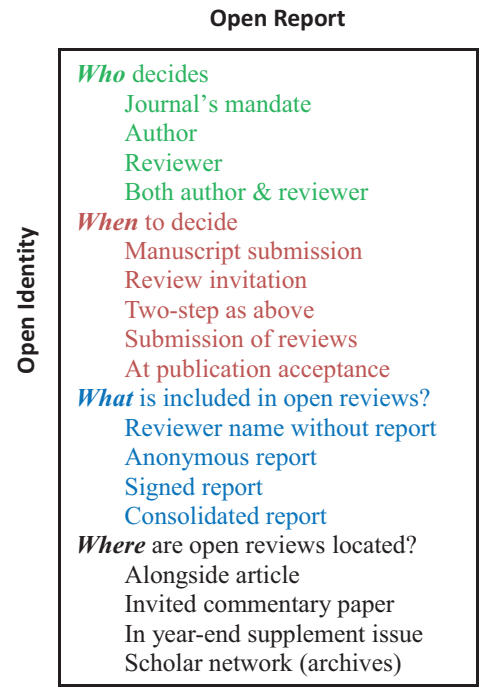


Figure 3. Factors in OPR Implementations

Discussion

This study represents the first comprehensive investigation of the scope and depth of OPR adoption in the open science era. Since the *BMJ* experiments with open reviews more than 20 years ago, the adoption of OPR has gone from 38 journals in 2001, to at least 174 journals by the end of 2018. Figure 1 demonstrates that there has been steady growth in the number of OPR journals over time, led by journals in medicine and the natural sciences. The remaining disciplines have been much slower and later to adopt OPR, especially the humanities. The

humanities have different scholarship cultures as compared to the natural sciences and have been slow in adopting open access (Eve, 2017; Gross and Ryan, 2015).

Several publishers have served as pioneers and early adopters of OPR. The three most prolific publishers of OPR journals that have led the way--BioMed Central, Frontiers, and Copernicus Publications--have each adopted different approaches. BioMed Central, as the leading OPR journal publisher in this study, began the practice early with dozens of journals, opting for both open reports and open identities. The publishers of 170 out of the 174 OPR journals in this study are based in Europe, signifying Europe's leading role in the OPR movement. The European scientific communities have been strong innovators in open science, so it is no surprise that European publishers would be innovators and early adopters in OPR. Three of the remaining journals are associated with publishers in the United States and one is published in Argentina. This strong European effort is also seen in the larger open science movement, where organizations such as OpenAIRE and OpenUP are investigating all aspects of this movement, including OPR.

Multiple OPR models emerge from the analysis of the data that show different levels of transparency in implementation. The level of transparency can be characterized along a continuum; a scoring algorithm is being developed and tested to compare different models using the process-product model incorporating the four factors (open identity, open report, what included, where to access) and process and product. The most transparent model is the concurrent open review process exemplified by *F1000Research*, where reviewers' identities and reports are instantly available alongside manuscripts and are published upon submission following initial format checking. Another model that promotes total transparency, exemplified by many BioMed Central journals, provides access to the complete report history and author exchanges as well as open identities alongside the published articles. The next several models that allow authors and/or reviewers to participate in open review decisions during the process include: mandated open reports but optional open identities (e.g., *eLife*), mandated open reports without open identities (e.g., *Ledger*), and optional open reports with optional open identities (e.g., *PeerJ*). The least transparent model, used by the Frontiers journals, is a closed review process with the published articles including only the names of the reviewers.

Conclusion

The adoption of OPR innovation is growing. This growth has been largely spurred by three publishers based in Europe. To date, OPR has been adopted mostly by journals in medicine and the natural sciences, although the number of OPR journals remains a very small percentage of scholarly journals, overall. The fact that there are multiple approaches to the adoption of OPR indicates there is no consensus at present regarding best practices. The gold standard for OPR transparency includes open identities along with open reports, but few OPR journals have adopted complete transparency.

Limitations of the present research must be recognized. Currently, there is no universal way to identify journals that adopt OPR models. Our approach was to cast a broad net using multiple sources to identify candidate journals. It is possible that we have missed OPR journals that are not indexed by sources such as DOAJ or the search services used. Like any indexing source, there may also be a regional or language bias. Also, the coverage of multidisciplinary journals may span more than one of the identified disciplines. These journals were categorized into the most relevant discipline.

The next phase of this research, currently underway, is analysing the contents of open reports under different models using text mining and natural language processing techniques to determine if the referee comments and quality differ under different models that support open reports and open identities.

Acknowledgments

This research was funded by a University of Wisconsin-Milwaukee Research Growth Initiative Grant. Peiling Wang acknowledges travel support from the University of Tennessee, Knoxville.

References

Barbash, F. (2014). Scholarly journal retracts 60 articles, smashes 'peer review ring', *The Washington Post*, July 10, https://www.washingtonpost.com/news/morning-mix/wp/2014/07/10/scholarly-journal-retracts-60-articles-smashes-peer-review-ring/?utm_term=.4ab26f14adb9

Belluz, J., Plumer, B., & Resnick, B. (2016). The 7 biggest problems facing science, according to 270 scientists. *Vox* http://www.vox.com/2016/7/14/12016710/science-challeges-research-funding-peer-review-process

Bohannon, J. (2013). Who's afraid of peer review? A spoof paper concocted by *Science* reveals little or no scrutiny at many open-access journals. *Science*, 342, 60–65. https://www.sciencemag.org/content/342/6154/60.full.pdf

Bornmann, L., Wolf, M., & Daniel, H. D. (2012) Closed versus open reviewing of journal manuscripts: how far do comments differ in language use? *Scientometrics*, 91, 843–856 DOI 10.1007/s11192-011-0569-5

Eve, M. P. (2017). Open access publishing models and how OA can work in the humanities. *Bulletin of the Association for Information Science & Technology*, 43(5), 16-20.

Ford, E. (2013). Defining and Characterizing Open Peer Review: A Review of the Literature. *Journal of Scholarly Publishing*, 44(4), 311-326.

Fresco-Santalla, A. & Hernández-Pérez, T. (2014). Current and evolving models of peer review. The Serials Librarian, 67:4, 373-398. doi: 10.1080/0361526X.2014.985415

Lee, C. J., Sugimoto, C. R., Zhang, C., & Cronin, B. (2013). Bias in peer review. *Journal of the American Society for Information Science and Technology*, 64 (1), 2-17

Gross, J, & Ryan, J. C. (2015) Landscapes of research: perceptions of open access (OA) publishing in the arts and humanities. *Publications*, 3, 65-88.

Godlee, F., Gale, C. R., & Martyn C. N. (1998). Effect on the quality of peer review of blinding reviewers and asking them to sign their reports: a randomized controlled trial. *JAMA,* 280, 237–40

Görögh, E., Schmidt, B., Banelytė, V., Stanciauskas, V., & Woutersen-Windhouwer, S. (2017). OpenUP Deliverable D3.1– Practices, evaluation and mapping: Methods, tools and user needs. OpenUP Project. http://openup-h2020.eu/wp-content/uploads/2017/01/OpenUP_D3.1_Peer-review-landscape-report.pdf

Kriegeskorte, N. (2012). Open evaluation: a vision for entirely transparent post-publication peer review and rating for science. *Frontiers in Computational Neuroscience*, 6, Article 79, 2-18

Mehmani, B. (2016). Is open peer review the way forward? https://www.elsevier.com/reviewers-update/story/innovation-in-publishing/is-open-peer-review-the-way-forward

Mehmani, B. & van Rossum, J. (2015). Elsevier trials publishing peer review reports as articles https://www.elsevier.com/reviewers-update/story/peer-review/elsevier-pilot-trials-publishing-peer-review-reports-as-articles

Mulligan, A., Hall, L., & Raphael, E. (2013). Peer review in a changing world: An international study measuring the attitudes of researchers. *Journal of American Society for Information Science and Technology*, 64: 132–161. doi: 10.1002/asi.22798

Peer Review Congress (2017). Under the Microscope: Transparency in Peer Review. Panel after the Peer Review Congress. Peer Review Congress, Chicago, 10-12, September, 2017. Panel chaired by Alice Meadows (ORCID) with panellists: Irene Hames (Board member of Learned Publishing), Elizabeth Moylan (BMC), Andrew Preston (Publons), and Carly Strasser (Moore Foundation) https://peerreviewweek.files.wordpress.com/2017/05/prw2017-panelists22.pdf Video at https://www.youtube.com/watch?v=8x1dho6HRzE

Pöschl, U. & Koop, T. (2008). Interactive open access publishing and collaborative peer review for improved scientific communication and quality assurance. *Information Services & Use*, 28, 105-7

Rath, M. & Wang, P. (2017). Open peer review in the era of open science: A pilot study of researchers' perceptions (Poster Paper) In *Proceedings of the Joint Conference on Digital Libraries (JCDL)*, 317-8

Ross-Hellauer T. (2017). What is open peer review? A systematic review. *F1000Research*. 2017 apr; 6:588. Available from: https://doi.org/10.12688%2Ff1000research.11369.1. PMID: 28580134
Ross-Hellauer, T., Deppe, A., & Schmidt, B. (2017). Survey on open peer review: Attitudes and experience amongst editors, authors and reviewers. *PLoS ONE,* 12(12), e0189311 https://doi.org/10.1371/journal.pone.0189311
Smith, R. (2006). Peer review: a flawed process at the heart of science and journals. *Journal of the Royal Society of Medicine*, 99, 178-182
Tattersall, T. (2015). For what it's worth – the open peer review landscape, *Online Information Review*, 39(5), 649-663, https://doi.org/10.1108/OIR-06-2015-0182
Tennant, J. P., Dugan J. M., Graziotin D., Jacques, D. C., Waldner, F., Mietchen, D., … Colomb, J. (2017). A multi-disciplinary perspective on emergent and future innovations in peer review. F1000Research 2017, 6:1151 https://doi.org/10.12688/f1000research.12037.3
van Rooyen, S., Godlee, F., Evans, S., Black, N., & Smith, R. (1999). Effect of open peer review on quality of reviews and on reviewers' recommendations: a randomised trial. BMJ, 318, 23–7.
Vrana, R. (2017). Acceptance of Open Peer Review by Scientific Publishers. *The 28th International Central European Conference on Information and Intelligent Systems*, September 27-29, Varazdin, Croatia. https://www.researchgate.net/publication/322919350_Acceptance_of_Open_Peer_Review_by_Scientific_Publishers
Wang, P. & Tahamtan, I. (2017). The State-of-the-Art of Open Peer Review: Early Adopters. (Poster Paper) *Proceedings of the 2017 Annual Meeting of The Association for Information Science & Technology*, October 27 – November 1, Washington DC.
Wang, P. & Wolfram, D. (2016). The Last Frontier in Open Science: Will Open Peer Review Transform Scientific and Scholarly Publishing? at the 2016 Annual Meeting of the Association for Information Science and Technology Oct. 14-18, 2016, Copenhagen, Denmark. [Panellists: Jason Hoyt, PeerJ; Ulrich Pöschl, Max Planck; Peter Ingwersen, Royal School of Denmark & Richard Smith, retired editor of The BMJ; discussant: Marcia Bates, University of California, Log Angeles]
Wang, P., You, S., Rath, M., & Wolfram, D. (2016). Open Peer Review in Scientific Publishing: A Web Mining Study of PeerJ Authors and Reviewers*, Journal of Data and Information Science* 1 (4), 60 – 80. https://content.sciendo.com/view/journals/jdis/1/4/article-p60.xml

Appendix – OPR Journals

'*' indicates an early OPR adopter from 2001

BioMed Central

Archives of Public Health; BioData Mining; Biology Direct; BMC Anesthesiology; BMC Cancer*; BMC Cardiovascular Disorders*; BMC Clinical Pathology*; BMC Complementary and Alternative Medicine*; BMC Dermatology*; BMC Ear, Nose and Throat Disorders*; BMC Emergency Medicine*; BMC Endocrine Disorders*; BMC Family Practice*; BMC Gastroenterology*; BMC Geriatrics*; BMC Health Services Research*; BMC Hematology** (prev. *BMC Blood Disorders*)*; BMC Infectious Diseases*; BMC International Health and Human Rights*; BMC Medical Education*; BMC Medical Ethics*; BMC Medical Genetics*; BMC Medical Genomics*; BMC Medical Imaging*; BMC Medical Informatics and Decision Making*; BMC Medical Research Methodology*; BMC Medicine; BMC Musculoskeletal Disorders*; BMC Nephrology*; BMC Neurology*; BMC Nursing*; BMC Nutrition; BMC Obesity; BMC Ophthalmology*; BMC Oral Health*; BMC Palliative Care*; BMC Pediatrics*; BMC Pharmacology and Toxicology; BMC Pregnancy and Childbirth*; BMC Psychiatry*; BMC Psychology; BMC Public Health*; BMC Pulmonary Medicine*; BMC Rheumatology; BMC Sports Science, Medicine and Rehabilitation; BMC Surgery*; BMC Urology*; BMC Women's Health*; Cardiovascular Ultrasound; Diagnostic and Prognostic Research; Environmental Health; Head & Face Medicine; Health Research Policy and Systems; Hereditary Cancer in Clinical Practice; Human Resources for Health; Implementation Science; Journal of Cardiothoracic Surgery; Journal of Foot and Ankle Research; Journal of Medical Case Reports; Nutrition Journal; Pilot and Feasibility Studies; Population Health Metrics; Reproductive Health; Research Integrity and Peer Review; Research*

Involvement and Engagement; Scoliosis and Spinal Disorders (prev. *Scoliosis); Systematic Reviews; Trials*

Copernicus Publications
Annales Geophysicae; Atmospheric Chemistry and Physics; Atmospheric Measurement Techniques; Biogeosciences; Climate of the Past; Drinking Water Engineering and Science; Earth Surface Dynamics; Earth System Dynamics; Earth System Science Data (ESSD); Geoscience Communication; Geoscientific Instrumentation, Methods and Data System; Geoscientific Model Development; Hydrology and Earth System Sciences; Natural Hazards and Earth System Sciences; Nonlinear Processes in Geophysics; Ocean Science; SOIL; Solid Earth; The Cryosphere; Wind Energy Science*

Elsevier
Agricultural and Forest Meteorology; Annals of medicine and surgery; Engineering Fracture Mechanics; International Journal of Surgery; Journal of Hydrology: Regional Studies

Frontiers
Frontiers for Young Minds; Frontiers in Aging Neuroscience; Frontiers in Applied Mathematics and Statistics; Frontiers in Artificial Intelligence; Frontiers in Astronomy and Space Sciences; Frontiers in Behavioral Neuroscience; Frontiers in Big Data; Frontiers in Bioengineering and Biotechnology; Frontiers in Blockchain; Frontiers in Built Environment; Frontiers in Cardiovascular Medicine; Frontiers in Cell and Developmental Biology; Frontiers in Cellular and Infection Microbiology; Frontiers in Cellular Neuroscience; Frontiers in Chemistry; Frontiers in Communication; Frontiers in Computational Neuroscience; Frontiers in Digital Humanities; Frontiers in Earth Science; Frontiers in Ecology and Evolution; Frontiers in Education; Frontiers in Endocrinology; Frontiers in Energy Research; Frontiers in Environmental Science; Frontiers in Evolutionary Neuroscience; Frontiers in Forests and Global Change; Frontiers in Genetics; Frontiers in Human Neuroscience; Frontiers in ICT; Frontiers in Immunology; Frontiers in Integrative Neuroscience; Frontiers in Marine Science; Frontiers in Materials; Frontiers in Mechanical Engineering; Frontiers in Medicine; Frontiers in Microbiology; Frontiers in Molecular Biosciences; Frontiers in Molecular Neuroscience; Frontiers in Neural Circuits; Frontiers in Neuroanatomy; Frontiers in Neuroenergetics; Frontiers in Neuroengineering; Frontiers in Neuroinformatics; Frontiers in Neurology; Frontiers in Neurorobotics; Frontiers in Neuroscience; Frontiers in Nutrition; Frontiers in Oncology; Frontiers in Pediatrics; Frontiers in Pharmacology; Frontiers in Physics; Frontiers in Physiology; Frontiers in Plant Science; Frontiers in Psychiatry; Frontiers in Psychology; Frontiers in Public Health; Frontiers in Research Metrics and Analytics; Frontiers in Robotics and AI; Frontiers in Sociology; Frontiers in Surgery; Frontiers in Sustainable Food Systems; Frontiers in Synaptic Neuroscience; Frontiers in Systems Neuroscience; Frontiers in Veterinary Science

F1000Research Ltd.
F1000Research; Gates Open Research

Other Publishers
CVIR Endovascular; Economics: The Open-Access, Open-Assessment e-Journal; eLife; European Cells & Materials; GigaScience; Journal of Open Psychology Data; Journal of Open Source Software; Ledger; Papers in Physics; PeerJ; Research Ideas and Outcomes; Royal Society Open Science; SciPost Physics; Webmed Central; Wellcome Open Research*

Scientometric method for Comparing on the Performance of Research Units in the Field of Quantum Information

Yunwei CHEN[1] Zhiqiang ZHANG[1*] Cheng TAO[2] Jing XU[1] Qianfei TIAN[1] Jorge GULÍN-GONZÁLEZ[3] and Qiang LIU[1,4]

[1] *chenyw@clas.ac.cn, zhangzq@clas.ac.cn, jingxu@clas.ac.cn, tqf@clas.ac.cn*
Scientometrics & Evaluation Research Center (SERC), Chengdu Library and Information Center of Chinese Academy of Sciences, Chengdu, 610041 (China)

[2] *ctao@cashq.ac.cn*
Bureau of Development and Planning, Chinese Academy of Sciences, Beijing, 100864 (China)

[3] *Jorge.Gulin@neuroinformatics-collaboratory.org*
The Clinical Hospital of Chengdu Brain Science Institute, MOE Key Lab for Neuroinformation, University of Electronic Science and Technology of China, Chengdu, China and Universidad de las Ciencias Informáticas (UCI), Km 2½ Autopista a San Antonio de los Baños, La Habana, Cuba

[4] *liuqiang@mail.las.ac.cn,*
University of Chinese Academy of Sciences, Beijing, 100190 (China)

Abstract

Evaluation of countries, institutions and other social structures using scientific methods is a significant tool of governments and international organizations due to the impact in the decision making about public politics. Particularly this become critical for Research and Development systems at different levels (countries, institutions and others). A Research Unit (RU) consists of a relatively small group of scientists and researchers working in a particular research area. Usually, this structure is part of university or an institute. Here, we compare the Research Units (RUs) performance, based on scientometrics methods. Our methodology consists of four aspects: strategic vision, number of articles and citations, team structures represented by the scientists' average number of citations per article, ages and academic background of the scientists, and research modes represented by collaboration networks. This methodology has been used to compare the performance of four leading RUs in the field of quantum information, which were: CAS Quantum Center of Excellence (CAS-QC), MIT Center for Limit Quantum Information Theory (MIT-QIT), University of Oxford's Quantum Computing Center (Oxford-Q), and Institute of Quantum Science and Technology, University of Calgary (UC-IQST). Our attention was focused on the CAS-QC. Taking into account the obtained results, we concluded that each RU showed strengths and weaknesses. Particularly, CAS-QC showed a lower number of scientific articles, a lower share of highly cited scientists and a weaker variety of disciplines comparing to the MIT-QIT and the Oxford-Q. In contrast, CAS-QC had a higher number of young scientists, it is an obvious strategic advantage for competition. Finally, CAS-QC has a solid team structure and a united research paradigm which is a better way to exploit team superiority and tackle scientific challenges. The obtained results in this study are a validation of the propose methodology to investigate systems with similar characteristics.

Introduction

The work of evaluation has attracted significant attention in the last years by the impact in the decision making about public politics in economy, education, science and technology (Halevi, et al., 2017; Thelwall, 2017). Particularly evaluation become critical for Research and Development (R&D) systems. In general, the scientific evaluation can be performed by three different ways: using qualitative and quantitative methods or a combination of them. For quantitative evaluation of R&D entities, the scientometrics methods have been extensively used. The objects under study can be divided into three levels considering the dimensions of them: macro (countries), meso (institutions) and micro (authors or scientists). In the scientific literature are available many indexes, indicators and methods to carry out evaluation of case studies in several research fields. Country level evaluation analysis usually is focused on the performance of countries. For example, Prathap (2018) used matrix normalization and

multiplication to obtain the total input and output measures and carry out comparative research evaluation on countries levels. Concerning to the meso level evaluation, there are a lot of validated rankings of universities, such as QS World University Rankings, Times Higher Education World University Rankings, etc. In order to perform these studies many automatic tools are accessible on Internet, for example, Mingers *et al*. (2017) employed Google Scholar (GS) institutional level data to evaluate universities in a relatively automatic way. At the level of author, one of the most extended indicator is h-index, which is easy to use but with some shortages and argues.

In addition to the above three research levels, there is a very important type of scientific entity: The Research Unit (RU), which consists of a group of scientists and researchers. A research unit usually is a part of a university or an institute. It falls in between the meso and micro levels of research entities. In general, Research Units (RUs) are focused on specific research areas with relatively small scale. Some independent RUs gather to form colleges or institutes, which are major parts in universities. The performance of such RUs determines the university's success. Therefore, the competition between universities is essentially the competition between RUs consisting of a number of researchers focusing on a particular research field.

However, there are a few methods available to evaluate such kind of RUs. The purpose of the present study is to create a framework to compare the performance of RUs. As a case study was used the field of quantum information. Quantum information science and technology are based on fundamental physical principles. Quantum information science aims to develop new forms of information processing systems, spanning three broad categories: sensing, computing (including simulation), and networking (Monroe, 2019). The research and applications of quantum information technology might become one of the major breakthrough areas in the future. Many developed countries have devoted great importance to the R&D of quantum information and they have invested significant resources in order to seize the advantage of quantum information technology. According to the *Technical Opportunity for Quantum Era*, a report published by the UK government, the annual investment budget for quantum technology in six major countries/regions is no less than 100 million euros per region or country. Reported quantities are: 550 million euros in the EU, 360 million euros in the United States, 220 million euros in China, 120 million euros in Germany, 105 million euros in the United Kingdom and 100 million euros in Canada respectively (UK Government Office of Science, 2017). Particularly, China is investing great efforts to develop quantum communication in order to reach the leading position in the world and foster this as one of the representative national research fields. The relative achievements have attracted attention, especially since the launch of the Mozi quantum scientific experimental satellite in August 2016. Research has taken a leading position internationally. At present, the Mozi has successfully achieved all three established scientific goals (Yin *et al*., 2017, Liao *et al*., 2017, Ren *et al*, 2017). For China, this lays a solid scientific and technological base to continue leading the development of the world quantum communication technology and the frontier research on critical issues of quantum physics at the spatial scale. Since 2006, Mozi has realized an intercontinental quantum key distribution with a distance of 7,600 km between China and Austria and it has used shared keys for encrypted data transmission and video communication (Liao et al., 2018). The Beijing-Shanghai trunk line, the world's first 1000-km-level quantum security communication in this category, with a total length of more than 2,000 km, had also been put into use in 2017. These major achievements were accomplished by CAS Quantum Center of Excellence (CAS-QC). This team is responsible a series of major achievements in the field of quantum communication in China.

There are many studies based on articles or patents data to reveal the progress of quantum information. For example, Gao & Xu (2017) conducted basic quantitative statistics on the

articles and patents data in this field. Liu & Li (2014) used the method of co-word analysis to analyse the keywords of the Science Citation Index (SCI) articles from 2002 to 2011, and the result was used to discover the research hotspots of quantum information science. However, researches on the global competition for quantum information are still rare. In order to comprehensively reveals the current state of the international competition and quantum information patterns, we have used scientometric methods to study the global trends with the output of the articles. At the same time, we selected some leading international scientific research units to conduct comparative analysis that shown the current state of the research competition. The so-called scientific RU refers to research centers, key laboratories, centers of excellence, etc. established within scientific research institutes, universities, or R&D companies (Chen *et al*., 2015).

Data & Methods

The definition of quantum information stated by Qurope.eu website (Binosi & Calarco, 2017) was used for searching the total number of scientific articles about quantum information in the ISI-WOS database. All types of documents were retrieval on April 15, 2017.

The search strategy of Quantum Information was: TS=((Quantum and ((information) or (eraser) or (Quantum-Classical Transition) or (coherence) or (entanglement) or (measurement) or (network) or (storage) or (memory) or (communication) or (fingerprint) or (processor) or (Cavity QED) or (clock synchronization) or (imag*) or (sensor) or (magnetometry))) OR ((quantum NEAR/5 comput*) OR (quantum NEAR/15 algorithm*) OR (quantum NEAR/10 simulat*) OR (quantum NEAR/10 error*) OR (quantum circuit” OR “Quantum cellular automata” OR “Quantum Turing machine” OR “quantum register) OR (quantum NEAR/10 communication*) OR (quantum NEAR/15 protocol*) OR (quantum NEAR/15 cryptograph*) OR (“quantum key”))).

Using the above described method, TOP10 institutions with highest publication were obtained (Figure.1). After that, we selected the world's top RUs within these top 10 institutions with similar research fields and organizational models, and different research directions to carry out a comparative analysis. All publications of each scientist were downloaded and then aggregates them into a collection of articles for each research unit. The information about the academic background and the age of the scientists were consulted from the official website of each research unit or personal home page of the scientist. Thomson Data Analyzer (TDA) software was used for statistical analysis of the articles, and the scientific cooperation network analysis was performed using the Science of Science (Sci2) software (Sci2 team, 2009).

Methods Framework for Comparing the Performance of Research Units

Based on the idea of *Leiden Manifesto* (Hicks *et al*., 2015), with the purpose to evaluate research unit basing on a group of comprehensive methods, here we propose a new framework for comparing the performance of RUs, which consists of the following elements:

1. Strategic vision, an indicator to measure performance against the research missions of the research units.
2. Number of articles and citations, it is a traditional but essential indicator for evaluations.
3. Team structure, represented by scientist’s average times cited per article, ages and academic background, which is a new aspect for evaluation.
4. Research modes, represented by scientists’ collaboration network within a unit.

A case study on four research units in the field of quantum information

Through the statistics of global articles in quantum information, we got the list of TOP10 institutions with highest publication in this field (Figure.1). Thus, taking into account the expert opinions, a sample of world's top research units with similar research fields and

organizational models and different research directions were compared. The research units include the CAS Quantum Center of Excellence that is at the leading position in China (referred to as the “CAS-QC”), the MIT Center for Limit Quantum Information Theory (referred to as the “MIT-QIT”) that is the world leader in quantum theory research and University of Oxford's Quantum Computing Center (referred to as "Oxford-Q") that is unique in quantum computing and applied research was conducted. In addition, we also observed that the University of Calgary, Canada, established the Institute of Quantum Science and Technology (referred to as "UC-IQST") in 2013, focusing on frontier topics of the quantum information theory and experimental research, has also achieved international leading results in the use of urban fiber networks to achieve quantum communication research (Science and Technology Daily, 2016). We also included the new team in the comparative analysis. Finally, we chose the three research units to conduct comparative analysis with the CAS-QC.

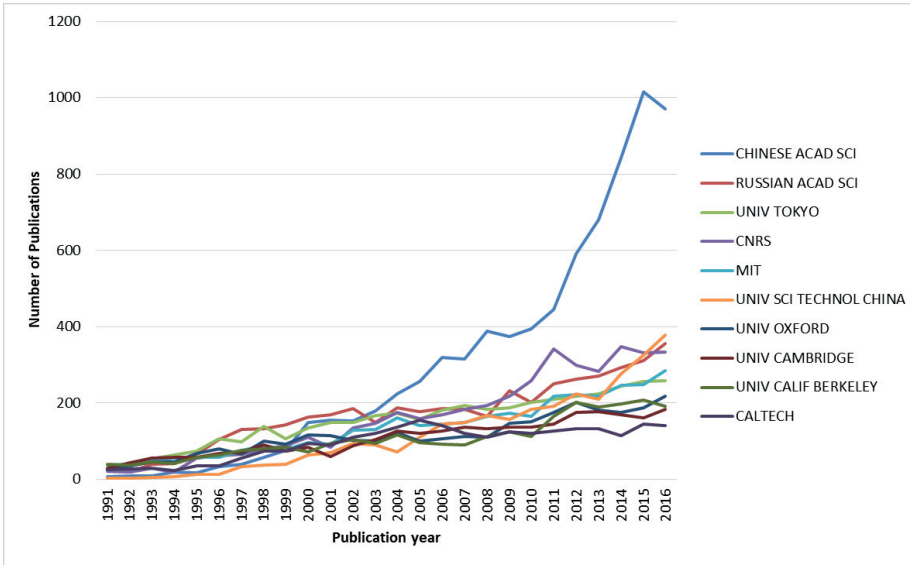


Figure 1 TOP 10 institutions with highest publication in quantum information.

Under the background of the current global competition of quantum information research, China's quantum information research has a leading edge in the world. In the following section, the comparative analysis of China’s leading research unit CAS-QC and the internationally leading scientific research unit MIT-QIT, Oxford-Q and UC-IQST will be conducted to observe current international competition in quantum communications and quantum computing.

Strategic vision and basic information of each research unit

The mission of CAS-QC is to carry out experimental and applied researches on quantum communication and quantum computing. The mission of MIT-QIT is to explore and solve three major theoretical issues: the NP complete problem in the extreme quantum information environment, quantum communication coding technology, and the basic physical limitations of quantum sensing and control. On the other hand, the mission of the Oxford-Q is to use a new generation of devices to better explore quantum utilities. Finally, the mission of the UC-IQST is to conduct the frontier researches on the key theoretical and experimental topics in quantum science and technology (Table 1).

Table 1 Introduction of four RUs in the field of quantum information

Units	*Country*	*Found year*	*Goals and vision*	*Core Scientists**
MIT-QIT	USA	2007	Explore and solve three major theoretical issues.	9
Oxford-Q	UK	1998	For quantum technology, use a new generation of devices to better explore quantum utility.	45
CAS-QC	China	2017	Experimental and applied researches on quantum communication and quantum computing.	63
UC-IQST	Canada	2013	Conduct the frontier researches on the key theoretical and experimental topics in quantum science and technology	14

(*Notes: By the end of April 15, 2017.)

Articles and citations

By counting the SCI articles of the four studied RUs based on the total scientific production of scientist (as of April 15, 2017), we found that the number of articles of CAS-QC and Oxford-Q were much more than the other two RUs (Figure 2). However, in terms of the number of articles per scientist, the MIT-QIT has the highest value both on articles and citations per scientist.

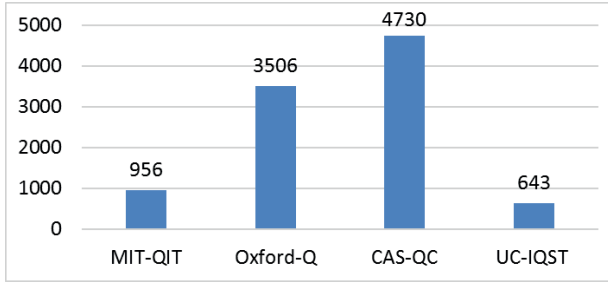


Figure 2 Number of articles per RU.

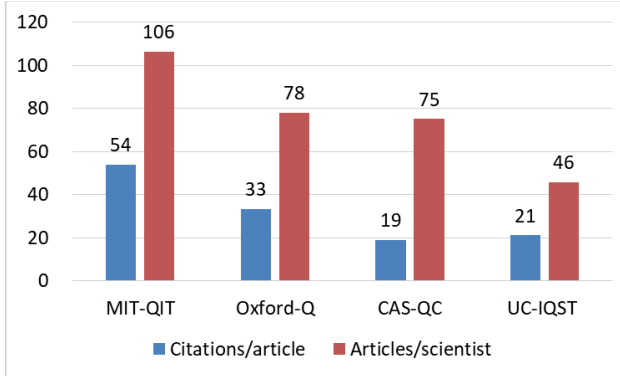


Figure 3 Number of articles and citations per scientist in the studied RUs

In addition, this study also counts the number of articles published on the two high level international journals of Science and Nature by the four research units. In this aspect, Oxford-

Q has an obvious advantage with respect to the other three units. The number of articles published in the two journals by CAS-QC is as much as that of MIT-QIT (Figure 4).

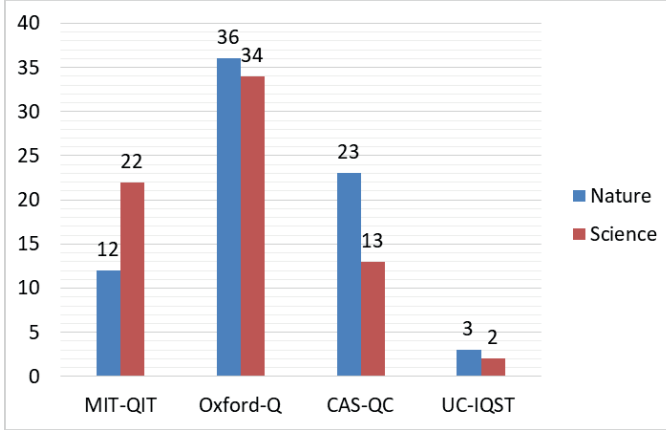


Figure 4 Articles published on the top journals of Science and Nature in the studied RU.

Team Structures

Our study examined the team structure of each unit from the following three perspectives: the distribution of scientists based on their articles' average times cited, age, and academic background.

First, scientists' distribution pyramid of each unit was constructed based on the a scientist's average number of citations per article (Figure 5). It can be seen that the values of MIT-QIT were consistently distributed from high to low (with prevalence to high), which means that MIT-QIT has a good personnel structure with a high research impact. While most scientists at the Oxford-Q and CAS-QC were at the bottom of the pyramid, which means that most scientists at the two units have weak research impacts counted by average number of citations of each scientist's articles. Particularly, the scientists of CAS-QC had a higher rate in the bottom of the pyramid. In comparison, the UC-IQST had the lowest ratio of high-level scientists and the highest ratio of low academic influence scientists.

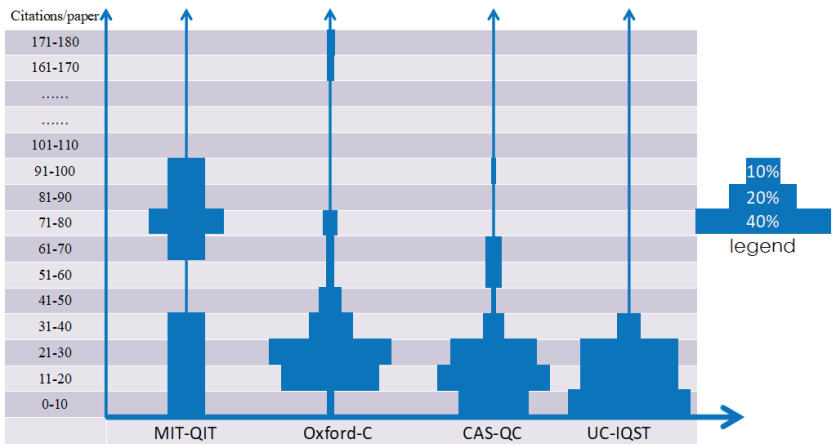


Figure 5 Pyramid distribution of scientists based on times cited per article in the studied RUs

Second, CAS-QC had the youngest team (highest ratio of young scientists), while the other three research units had percentage of middle-aged scientists (Figure 6). The average age of the scientists in CAS-QC was 43.8 years, while the average ages of scientists at the MIT-QIT, Oxford-Q and UC-IQST was 62.2 years, 49.9 years and 49.6 years, respectively. Most of the scientists in the CAS-QC had 30 to 39 years old (40%) and 40 to 49 years old (40%). Most of the scientists in the MIT-QIT had 50 to 59 years old (44%). Most of the scientists in the Oxford-Q and UC-IQST had 40 to 49 years old, accounting for 40% and 50% respectively. According to some studies about the peak age in that the Nobel Prize Winners' produce the representative work (Men & Zhang, 2013; Bjork, 2019), we can conclude that the CAS-QC scientists are at the optimal stage for the science and major knowledge innovations.

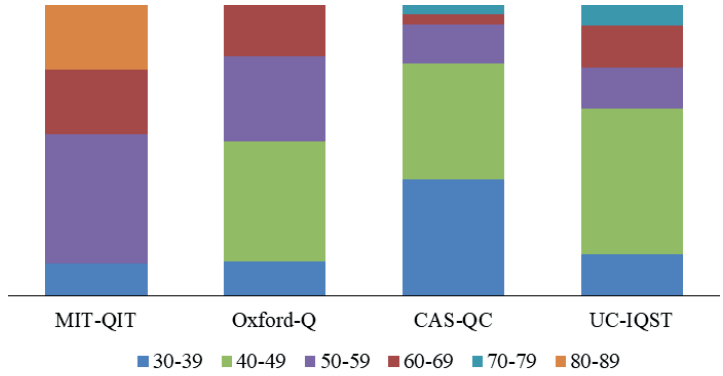


Figure 6 Age distribution of scientists in the studied RUs

Third, the academic background of the scientists in CAS-QC focus on the field of physics, while the academic backgrounds of the other three RUs were more diversified. Particularly, the Oxford-Q has the most extensive professional background in personnel disciplines and even includes scientists with theological philosophy and theology background (Figure 7).

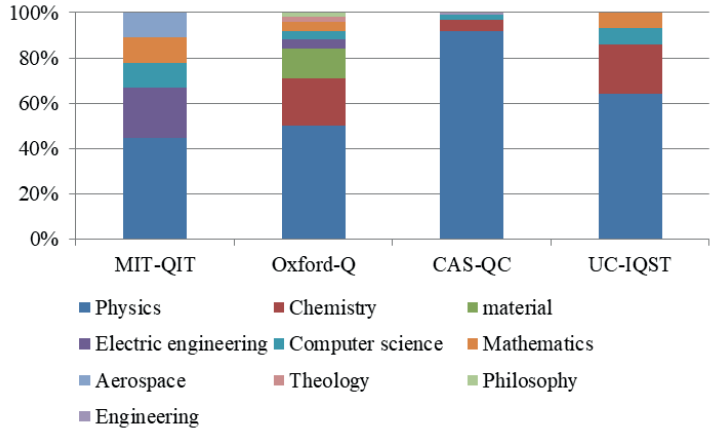


Figure 7 Academic background of the scientists in the studied RUs

Comparison of Scientific Research Modes

The structures of scientist collaboration networks had been used to reveal the characteristics of the four RUs, which is called scientific research modes in this paper referring to the disciplinary structure of the scientific units and the behavioral characteristics. It is feature for distinguishing the research characteristics of RUs and can be revealed by the scientists' collaborative networks (Chen et al., 2016).

CAS-QC presented a strong group-style research mode and has a strong joint research feature, while the scientists of the MIT-QIT and Oxford-Q are relatively independent. As it shown in Fig 8(A), the CAS-QC has formed a number of intensive collaborative clusters.

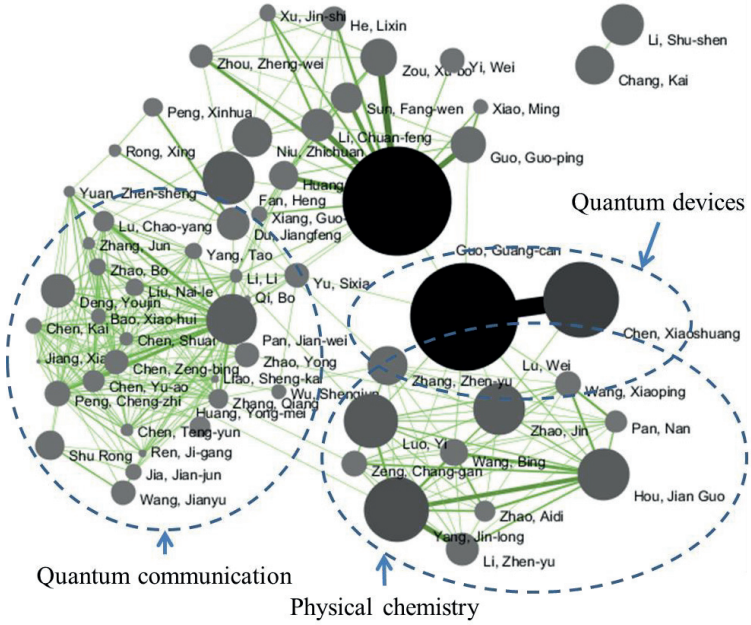


Figure 8 (A) Collaboration network of CAS-QC(only 63 scientists are shown)

The scientists' collaborative network of MIT-QIT was found relatively sparse (Figure 8(B)), with a certain degree of weak-strength collaborations among 9 scientists, who tend to conduct the research independently. Each of these scientists had a relatively independent collaboration network and their partners are mostly postdoctoral fellows, visiting scholars and students. Among the 9 core scientists, many of them are leaders or founding talents in the field of quantum information. For example, Jeffrey Goldstone proposed an important theorem in 1962 (Goldstone et al., 1962). He is one of the most important researchers of quantum field theory and he proved the Higgs boson theory. Peter Shor is the author of the quantum decomposition algorithm (Shor algorithm) (Shor, 1997), which is the most relevant in the field of quantum computing. He proved that quantum computers can be used to crack the RSA encryption algorithms that are currently used and set off the culmination of research on quantum computers. Researchers of CAS-QC used the photon bit experiment to implement the Shor algorithm for the first time in 2007 (Lu, et al., 2007). Researchers of MIT firstly implemented the Shor algorithm in a scalable manner in 2016 (Monz, et al., 2016).

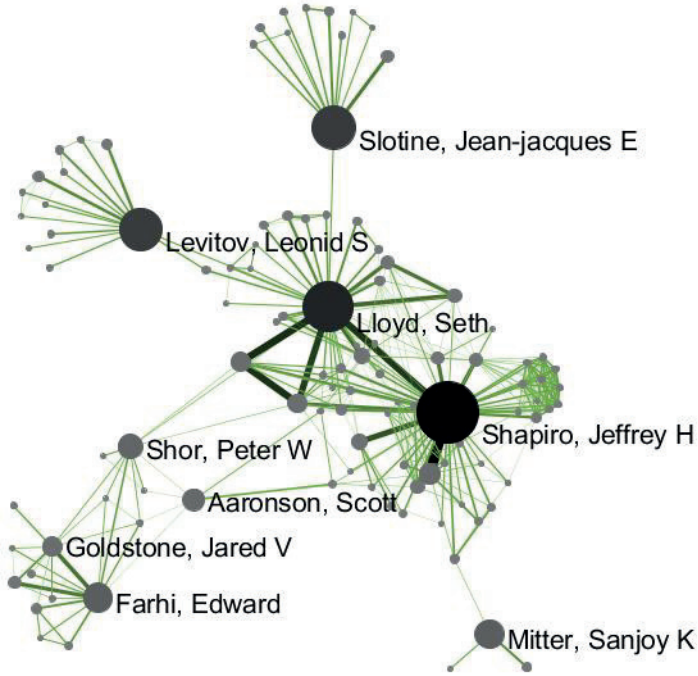


Figure 8 (B) Collaboration network of MIT-QIT. Here, only scientists with more than 5 articles are shown

Among the 45 scientists of the Oxford -Q (Figure 8(C)) in the field of quantum materials showed a relatively high degree of collaboration. Most of the collaborations in other field were within teams.

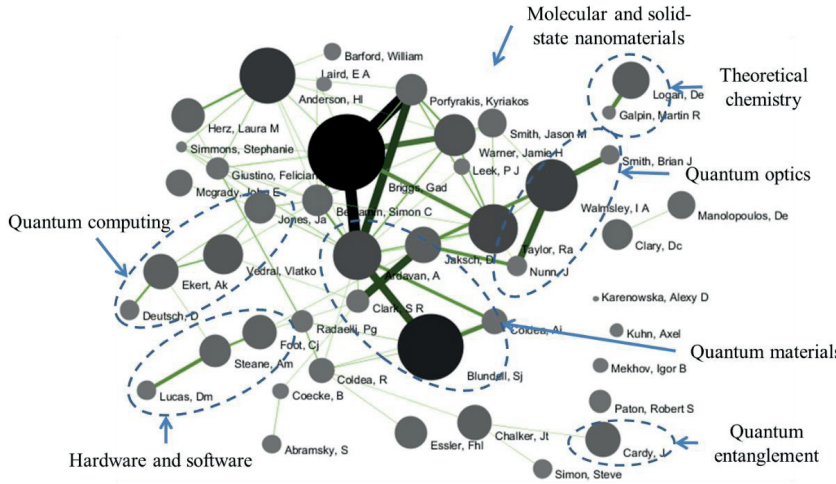


Figure 8 (C) Collaboration network of Oxford-Q. In this figure the most significant 45 scientists are shown

The research mode of UC-IQST falls in between CAS-QC and MIT-QIT or Oxford-Q. Some scientists form a close cluster and some are independent who have few collaborations to other (Figure 8(D)).

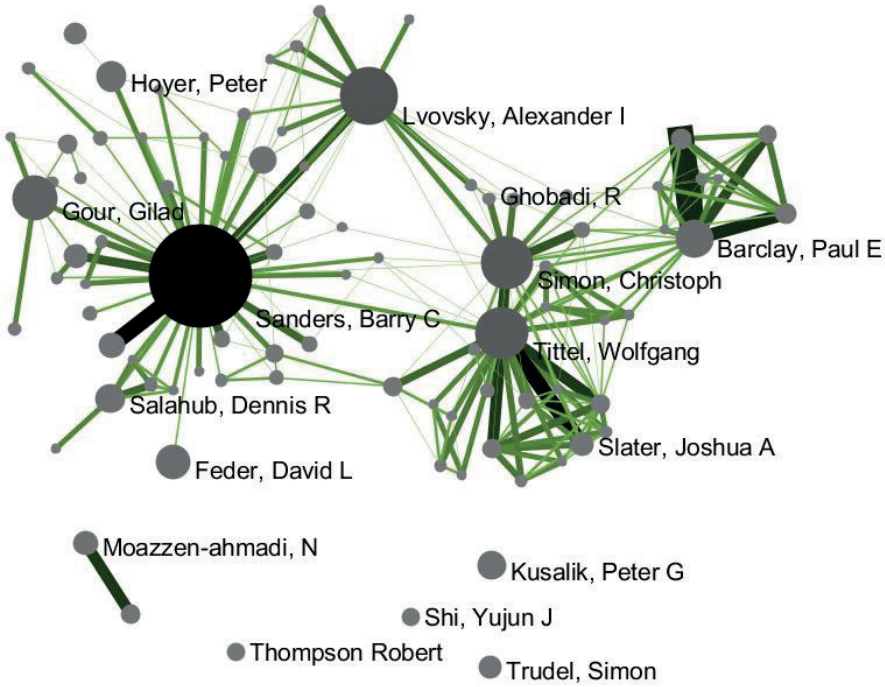


Figure 8 (D) Collaboration network of UC-IQST

Results and Discussion

This investigation compared the performance of four leading RUs engaged in the field of quantum information through a framework of scientometric methods. The result showed that the number of articles of CAS-QC was larger in comparison with the other three studied units but the overall impact of them was lower than MIT-QIT and Oxford-Q. Concerning to the team structure evaluated taking into account the highly cited scientists, CAS-QC does not showed advantage with respect to MIT-QIT and Oxford-Q. CAS-QC, lacks of leadership scientists. Also, the scientists' academic background of CAS-QC was relatively monotonous. In contrast, the advantage of CAS-QC was it had much more young scientists who are at the optimal ages of creativities and challenge.

Another potential advantage of CAS-QC might be its research mode with the particular feature of big group focusing on a few particular research fields. It has to be noted that this advantage is highly dependent on the aims of each units. For example, MIT-QIT has outstanding advantages in theoretical research and has gathered a number of leading scientists. However, CAS-QC is good at experimental research and focusing on the key areas, for example, the field of quantum information (our case study), which needs big teams to work hard to achieve their goal.

Thus, this kind of team structure might increase the global role of CAS-QC in the development of the quantum information science. CAS-QC should pay much more attention to train its young scientists who are its one big advantage and might also be the key for future

success. These young scientists are in the best age to achieve major knowledge and innovation. CAS-QC should increase the diversity of academic backgrounds of its staffs.

Finally, the methodology used in our study has demonstrated to be efficient for evaluating the performance of RUs at different scales. It overcomes the shortage of traditional evaluation indicators only basing on articles and citations.

Acknowledgments

This work is funded by the Special Project for Strategic Research and Decision Support System Construction of the Chinese Academy of Sciences (GHJ-ZLZX-2019-31).

References

Binosi, D. & Calarco, T. (2017). Quantum Information Classification Scheme – QICS. Retrieved May 1, 2017 from: http://qurope.eu/content/qics-book.

Bjork, R. (2019). The age at which Noble Prize research is conducted. *Scientometrics*, 119(2), 931-939.

Chen, Y.W., Deng, Y., Chen, F., et al. (2015). Evaluation Model and Empirical Study of Scientific Research Units Innovation. *Journal of Intelligence*, 34(5),59-65 (In Chinese).

Chen, Y.W., Xu, J., Wang, L.N., et al. (2016). Research paradigm of scientific research units based on network structure. *Library and Information Technology*, (5),53-61 (In Chinese).

Gao, F. & Xu, F. (2017).The Latest Development of Global Quantum Information Technology and Its Enlightenment to China. *China Science and Technology Forum*, 5(06),164-170.

Goldstone, J., Salam, A., Weinberg, S. (1962). Broken Symmetries. *Phys. Rev*, 127, 965-970.

Halevi, G., Moed, H., & Bar-Ilan, J. (2017). Suitability of Google Scholar as a source of scientific information and as a source of data for scientific evaluation-Review of the Literature. *Journal of Informetrics*, 11(3), 823-834.

Hicks, D., Wouters, P., Waltman, L., et al. (2015). The Leiden Manifesto for research metrics. *Nature*, 520(7548),429-31.

Liao, S.K., Cai, W.Q., Liu, W.Y., et al. (2017). Satellite-to-ground quantum key distribution. *Nature*, 549,43-47.

Liao, S.K., Cai, W.Q., Handsteiner, J., et al. (2018). Satellite-Relayed Intercontinental Quantum Network. *Phys. Rev. Lett*, 120(3),030501.

Liu, X.P. & Li, Z.X. (2014). Analysis of the Hot Topics of Quantum Information Science Based on Co-word Analysis. *Science Observation*, 9(5),13-22.

Lu, C.Y., Browne, D.E., Yang, T., et al. (2007). Demonstration of a compiled version of Shor's quantum factoring algorithm using photonic qubits. *Physical Review Letter*s, 99(25),250504.

Men, W.L. & Zhang, Z.Q. (2013). Research on the Changing Age of Peak Age in Scientific Research Creation: A Case Study of Nobel Prize Winners in Natural Science. *Science in Science*, 31(8), 1152-1159 (In Chinese).

Mingers, J., O'Hanley, J. R., Okunola, M. (2017). Using Google Scholar institutional level data to evaluate the quality of university research. *Scientometrics*, 113(3), 1627-1643.

Monroe, C., Raymer, M. G., & Taylor, J. (2019). QUANTUM INFORMATION The US National Quantum Initiative: From Act to action. *Science*, 364(6439), 440-442.

Monz, T., Nigg, D., Martinez, E.A., et al. (2016). Realization of a scalable Shor algorithm. *Science*, 351(6277):1068.

Prathap, G. (2018). Totalized input-output assessment of research productivity of nations using multi-dimensional input and output. *Scientometrics*, 115(1), 577-583.

Ren, J.G., Xu, P., Yong, H.L., et al. (2017). Ground-to-satellite quantum teleportation. *Nature*, 549:70-73.

Science and Technology Daily. (2016). Two experiments have proved that it is feasible to perform quantum teleportation through urban networks. Retrieved May 1, 2017 from: http://www.most.gov.cn/gnwkjdt/201609/t20160921_127809.htm.

Sci2 Team. (2009). Science of Science (Sci2) Tool. Indiana University and SciTech Strategies, https://sci2.cns.iu.edu.

Shor, P. W. (1997). Polynomial-time algorithms for prime factorization and discrete logarithms on a quantum computer. *Siam Journal on Computing*, 26(5), 1484-1509.

Thelwall, M. (2017). Three practical field normalised alternative indicator formulae for research evaluation. *Journal of Informetrics*, 11(1), 128-151.

UK Government Office of Science. (2017). The Quantum Age: technological opportunity. Retrieved May 1, 2017 from: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/564946/gs-16-18-quantum-technologies-report.pdf.

Yin, J., Cao, Y., Li, Y.H., et al. (2017). Satellite-based entanglement distribution over 1200 kilometers. *Science*, 356(6343),1180-1184.

The rivalry between Bernini and Borromini from a scientometric perspective

Martin Wieland and Juan Gorraiz

martin.wieland@univie.ac.at; juan.gorraiz@univie.ac.at
University of Vienna, Vienna University Library, Department for Bibliometrics and Publication Strategies, Boltzmanngasse 5, A-1090 Vienna (Austria)

Abstract

From a historical point of view, Rome and especially the University of La Sapienza, are closely linked to two geniuses of Baroque art: Bernini and Borromini. In this study, we analyze their rivalry from a scientometric perspective. This study also serves as a basis for exploring which data sources may be appropriate for broad impact assessment of individuals. We pay special attention to encyclopaedias and other databases or types of publications that are not normally used for this purpose. The results show that some sources such as Wikipedia are not exploited according to the possibilities they offer, especially those related to different languages and cultures. Moreover, analyses are often reduced to a minimum number of indicators, which can distort the relevance of the outcome. Other sources normally not considered for this purpose, like JSTOR, PQDT, Google Scholar, etc. can provide more relevant or more abundant information than the typically used Web of Science Core Collection and Scopus. Finally, we also contrast opportunities and limitations of old and new (YouTube, Twitter) data sources (particularly the aspects quality and accuracy of the search methods). Much room for improvement has been identified in order to use data sources more accurately and efficiently.

Background

This year's ISSI conference is organized in Rome, the Eternal City, and is hosted by Sapienza University of Rome, one of the oldest universities founded in 1303 with the Papal bull. On this occasion, we present a study showing the rivalry between Gian Carlo Bernini and Francesco Borromini from a scientometric perspective who both competed each other primarily in Rome in the Baroque era. They were the leading architects in the first half of the seventeenth century and coined the Roman Baroque style. Their different characters, backgrounds and attitudes to life presumably provoked the antagonism and caused their rivalry (Burbaum, 1999; Morrissey, 2006). They were opponents in character, personality, artistic style and tastes. Borromini's rational geometry contrasts Bernini's emotional theatricality. Because of his devotion to San Carlo Borromeo, Borromini changed his real name, Francesco Castelli. Being cerebral more than sensual, silent by nature, celibate, deeply religious, dressed in funeral black like a Spaniard - only sported red garters and rosettes in his shoes - he frightened people and was an unhappy and pessimistic person, quarrelling even with his best patrons - including several popes - and closest friends (Blunt, 1979). He never amassed a large personal fortune, and in the culmination of one of his depressions, Borromini killed himself, literally falling on his own sword at age 68. He survived his own mortal blow for almost 24 hours, even managing to leave behind a first-hand account of his death. On the contrary, Gian Carlo Bernini was subtle, gracious, diplomatic, and moved easily through the courts of popes and princes. He was the definition of a childhood genius, recognized as a prodigy when he was only eight years old, and was appointed Chief Architect of Saint Peter's only at age 31. Bernini was a very sensual person. He had an affair with Costanza, the wife of one of his assistants, and when he suspected her of being involved with his brother Luigi, he badly beat him and ordered a servant to slash her face with a razor. While Constanza was jailed for adultery, Bernini was exonerated by Pope Urban VIII (Morrissey, 2006; Mormando, 2011). Subsequently, he practiced his faith more sincerely complying with popal orders, and at age 41 he wed a 22-year-old Roman woman, Caterina Tezio, in an arranged marriage. Caterina bore him eleven children, including youngest son Domenico, his first biographer (Mormando, 2011). Bernini remained physically and mentally vigorous and active until his death in Rome at age 81. He was buried with little public fanfare.

Curiously Borromini's and Bernini's paths converged at La Sapienza. During their early years, the two did not compete head-to-head, Bernini is rather known to have assisted Borromini in trying to obtain a post. In 1632 he wrote a recommendation letter for his former assistant for the position of architect of La Sapienza, Rome's university. Bernini's letter secured the papal appointment for Borromini and provided Borromini with the opportunity to create one of his great ecclesiastical masterpieces, Sant'Ivo alla Sapienza (Morrissey, 2006).
This historical background has triggered this study, which serves as a tribute to these two geniuses who somehow culturally mark this year's ISSI conference at the University of La Sapienza in Rome.

Introduction

The aim of this work is to estimate the resonance and prestige of these two geniuses several centuries after their death and to compare the footprints they have left in scholarly communication and beyond. A similar type of study has already been carried out in the past (Marx et al., 2011; Gorraiz et al., 2011&2015), which serves as a basis and inspiration. For this purpose, several types of bibliographic sources have been selected, which in turn are discussed in detail, both their capacity and their suitability to be applied for broader impact assessment.
In his most recent book Henk Moed (2017) describes which indicators can be used to measure different aspects of scientific communication in evaluative informetrics, and discusses their limitations as well as considerations that must be taken into account for a correct and responsible use. In this study we focus on data sources. We not only discuss the opportunities and limitations of old and new bibliographical and citation data sources, but also analyse their search features and suggest the most relevant information and indicators that can be extracted from them.
Among the sources considered are encyclopaedias, such as Wikipedia and Encyclopaedia Britannica, catalogues, and several databases including the ones that are commonly used for bibliometric analyses like Web of Science Core Collection and Scopus (Moed, 2017). The results obtained in other databases with a more disciplinary scope (e.g. JSTOR), or dealing with other publication types such as dissertations (e.g. PQDT) - a document type of high potential relevance and usually not considered in this kind of analyses –, or resulting in higher coverage like Google Scholar, Microsoft Academic and CrossRef all tend to provide more relevant or more abundant information than classis sources like WoS or Scopus (Sven et al. 2017; Moed, 2017). Last but not least we are also compare these results with the ones obtained from other web sources like Twitter and YouTube.
Some authors have already demonstrated the incomplete use of some data sources such as Wikipedia and the danger of using these data to compile new global and composite indicators as for example the Altmetric Attention Score (Gumpenberger et al., 2017). The incompleteness of data can seriously distort the obtained results and the significance of their interpretation. There is a bias towards the included sources, whereas missing sources are disadvantaged. Language, data availability, completeness and accuracy of the sources, and availability of indicators, are the issues to be tackled in this study.

Data sources and Methodology

A large amalgam of data sources is used in the study: encyclopaedias and bibliographic dictionaries, library catalogues, citation databases, subject-specific databases, and some alternative web sources. They are summarized in three groups below:

A) Encyclopaedias, Biographical Dictionaries and Reference Systems:

Wikipedia is our main data source among the encyclopaedias, because it is already a major source of the most common systems that trace new metrics and altmetrics (see

e.g. Altmetric.com or PlumX). Wikipedia is a multilingual, web-based, free encyclopaedia based on a model of openly editable and viewable content, a wiki (see https://en.wikipedia.org/wiki/Wikipedia). Time magazine stated that the open-door policy of allowing anyone to edit had made Wikipedia the biggest, most popular and possibly the best encyclopaedia in the worldOne of the characteristics of Wikipedia is the language diversity, which seems not having been used enough for scientometric applications so far. There are currently 301 language editions of Wikipedia (also called language versions, or simply Wikipedias). Fifteen of these have over one million articles each[1]. Actually, Wikipedia provides for each page very abundant and detailed information, comprising basic page information, page protection and properties, edit history, as well as page view statistics and WikiChecker (Katz and Rokach, 2017). In our study, we are focusing on following parameters or indicators: 1) Number of language editions. 2) Page length (in bytes), 3) Number of page watchers, 4) Number of page watchers who visited recent edits, 5) Number of redirects to this page, 6) Number of literature included (references, links, biographies, etc.), 7) Number of "What links here" (hyperlinks or web citations attracted), and 8) Number of page views and daily average. Analyses have been performed not only in the English edition, but also in other main languages, namely German, French, Italian, Russian and Spanish. The results were then compared. Furthermore, the Chinese edition has also been considered in order to provide a comparison with an emerging language.

According to prior studies comparing science articles from Wikipedia and Encycloplaedia Britannica (Giles, 2005), Wikipedia's level of accuracy approached that of Britannica. Therefore we also compared the results from Wikipedia English with the ones resulting from Encyclopædia Britannica online.

1. The Encyclopædia Britannica (Latin for "British Encyclopaedia"), is the oldest English-language encyclopaedia still in production In 1933, the Britannica became the first encyclopaedia to adopt "continuous revision", in which the encyclopaedia is continually reprinted, with every article updated on a schedule. In March 2012, Encyclopædia Britannica, Inc. announced it would no longer publish printed editions, and would focus instead on Encyclopædia Britannica Online.[2]
2. The World Biographical Information System (WBIS Online) is based on the digitization of K.G. Saur's microfiche editions of Biographical Archives. Each Biographical Archive covers a different language and cultural area. For the individual archives the original texts of biographical articles from various reference works were taken and facsimile reproductions made. It contains over 6 million people, 8.5 million original biographical articles, published from 1559 to the end of the 20th century. Currently 30 Biographical Archives are available in digitized form in WBIS Online.
3. Festschriften, an article or book honouring a respected person, are collected in IJBF ("Internationale Jahresbibliographie der Festschriften") through research in libraries and their bibliographic data are recorded. Since 1983, 809,000 articles from 36,000 Festschriften of the years 1977-2017 have been indexed.

B) Databases

The two main bibliometric data sources, Web of Science Core Collection and Scopus, are well-known within the scientometric community. Besides them, we have included following databases:

[1] "Wikipedia:List of Wikipedias". English Wikipedia. Retrieved January 5, 2019.

[2] https://en.wikipedia.org/wiki/Encyclop%C3%A6dia_Britannica

1. JSTOR provides access to more than 12 million academic journal articles, books, and primary sources in 75 disciplines (see https://about.jstor.org/).
2. ARTbibliographies Modern (ABM) provides full abstracts of journal articles, books, essays, exhibition catalogues, PhD dissertations, and exhibition reviews on all forms of modern and contemporary art.
3. ProQuest Dissertations & Theses Global (PQDT) enables the inclusion of theses and dissertations as important document types in our analyses (Andersen and Hammarfelt, 2011; Gorraiz et al., 2011 & 2015). This data source includes more than 2.7 million searchable citations to dissertation and theses from around the world from 1861 to the present day together with 1.2 million full text dissertations.

The results obtained from these three databases are compared with the ones resulting from WoS Core Collection (WoS CC) and Scopus. WoS CC included all the indexes (Proceedings, Books) and the Emerging Sources Citation Index since 2015. Furthermore, Google Scholar, Microsoft Academic and CrossRef have been consulted via Publish or Perish (Harzing, 2007).

C) Web Tools

Finally, YouTube, an American video-sharing website headquartered in San Bruno, California and now operating as one of Google's subsidiaries, and Twitter, an American online news and social networking service on which users post and interact with messages known as "tweets", have also been used in an explorative way.

All searches were carried out in December 2018 according to the search options and syntaxes available in each data source. Search strategy and manual disambiguation were similar to the procedures described in previous studies (Gorraiz, Gumpenberger and Wieland; 2011 & 2015). Searches were performed separately in title, descriptors and abstracts fields, as in the topic field including all these options and in full text if available. Different documents and publication types were differentiated. The search for documents either related to Borromini alone or for both artists together did not present many difficulties, since the name Borromini is not very common. However, the search concerning Bernini did pose many serious difficulties, since it is a fairly common Italian name. Manual disambiguation was necessary to clean the data. This was practiced as long as the number of retrieved items was not too high and allowed to do so. Otherwise, other approaches were used, such as excluding the publications of the authors with that name (as long as it was not a Bernini biographer like his son Diego) or refining the search to certain fields or topics and excluding those of the natural sciences. In these cases, normally related to the search in full text, we have opted for a compromise between recall and precision as discussed in detail for each data source in the results section.

Citation analyses were conducted using the “Cited reference search” feature in WoS Core Collection. The number of citing documents and all citations to documents containing Borromini and/or Bernini in the research fields “Work” or “Author” were retrieved. Although Bernini did not publish anything, his works and exhibits are cited under his author's name[3]. That is why we have included the two types of analysis in WoS CC. In Scopus the search was carried out with the help of the search option in "secondary documents".

Searches in Google Scholar, Microsoft Academic and CrossRef were performed using Harzing’s tool “Publish or Perish” using the most relevant meaningful search fields for each data source. Even more difficult were the searches in the sources YouTube and Twitter.

These tools do not indicate the total number of items retrieved, and as the waiting time increases, their number also increases until a “potential saturation” is reached. Due to the volatility and low reliability of the retrieved data, we have limited ourselves to the search for

[3] Borromini did publish only memoirs and notes, but also his architectural works are cited under his name.

both geniuses at the same time in this case, which is already very time consuming but reasonable. The results could then be checked manually in order to remove incorrect items.
The search difficulties and peculiarities found in each data source will also be discussed in detail in the next section.

Results

The results are presented in four groups according to the classification mentioned in the section "Data Sources and Methodology".

A) Encyclopaedias, Biographical Dictionaries and Reference Systems

Table 1 shows the top six language editions (also including the Chinese Edition) available for Bernini and Borromini in Wikipedia according to the number of views collected from 2015-07-01 to 2018-12-25. Furthermore, it contains information about the daily average of page views as well as the number of edits, editors and watchers. These data are directly available in Wikipedia under the feature "Langviews Analysis". Bernini is available in 66 languages, Borromini in 51 languages.

Table 1. Top language editions available in Wikipedia for both artists according the number of page views

Bernini

Ranking	Language	Page views	Daily average	Edits	Editors	Watchers
Total	66	3.867.188	3.035			
1	English	1.129.253	886	289	100	247
2	Italian	689.322	541	289	126	57
3	Spanish	531.767	417	87	62	47
4	Russian	256.726	202	36	25	unknown
5	French	196.362	154	79	47	42
6	German	164.713	129	48	41	38
12	Chinese	38.438	30	24	12	unknown

Borromini

Ranking	Language	Page views	Daily average	Edits	Editors	Watchers
Total	51	807.232	634			
1	Italian	262.248	206	170	84	unknown
2	English	157.593	124	33	25	59
3	Spanish	124.769	98	42	26	unknown
4	German	44.011	35	34	21	unknown
5	French	40.561	32	87	19	unknown
6	Russian	38.594	30	13	10	unknown
12	Chinese	5.458	4	5	5	unknown

The results show a higher degree of internationalisation for Bernini. He is more popular than Borromini according to the number of language editions (66 versus 51) as well as according to the numbers of page views. For Bernini, the most viewed edition is the English one, but not for Borromini, where the national interest seems to be higher as the international one.
Figures 1 to 2 show the trend of page views for both editions (English and Italian) between July 2015 and December 2018 for both artists. The highest number of page views of Bernini's English edition dates from December 2016. The only possible explanation we have found for

that maximum peak is Bernini's short film that premiered in that year (https://www.imdb.com/title/tt6289758/). This peak is also visible in Bernini's Italian edition for Bernini and also very clearly in Borromini's Italian Edition. The trends of both artists correspond quite well and corroborate that their stories are profoundly connected.

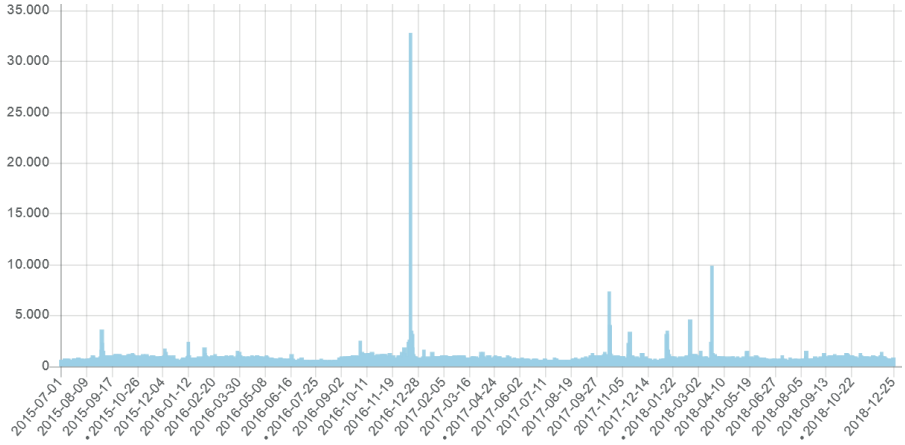


Figure 1. Page views of the English edition of Bernini between July 2015 and December 2018.

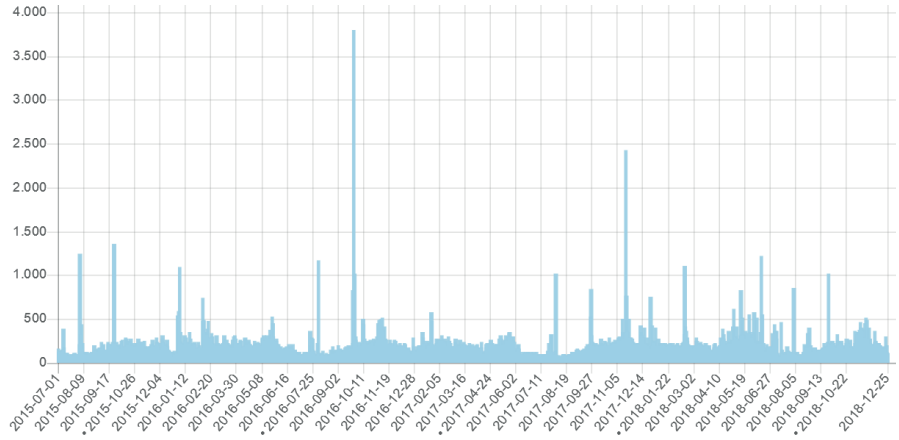


Figure 2. Page views of the Italian edition of Borromini between July 2015 and December 2018.

Table 2 gives an in-depth view of the Wikipedia results for both artists in the major six language editions (see section "Data Sources & Methodology") and the Chinese Edition.

Table 2. Short bibliometric analysis of the main language editions for each artist
Bernini (66 Languages)

	English	German	Italian	Spanish	French	Russian	Chinese
Page length (in bytes)	92,813	15,697	57,131	41,014	26,177	89,383	7,111
Number of page watchers	247	38	57	41	47	<30	<30
Number of page watchers who visited recent edits	25	8	13	10	9	n.a.	n.a.
Number of redirects to this page	10	3	4	3	4	11	4

References, Notes, Citations	77	17	29	9	10	66	4
Bibliographies	32	in RNC	25	7	15	30	n.a.
External Links	8	8	4	5	16	12	4
Honors (Memories)	n.a.	2	n.a.	n.a.	4	3	n.a.
Identifiers	17	5	10	19	n.a.	23	20
Categories	14	11	14	11	13	13	6
Pages that link to this one	1140	439	829	526	546	217	89
Borromini (51 Languages)							
	English	**German**	**Italian**	**Spanish**	**French**	**Russian**	**Chinese**
Page length (in bytes)	20,423	9,088	27,296	15,713	32,502	6,295	1,422
Number of page watchers	59	<30	<30	>30	<30	<30	<30
Number of page watchers who visited recent edits	6	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
Number of redirects to this page	7	2	2	1	1	4	1
References, Notes, Citations	23	8	13	6	28	1	1
Bibliographies	in RNC	in RNC	6	4	12	in RNC	in RNC
External Links	5	10	2	1	n.a.	4	n.a.
Honors (Memories)	2	1	1	n.a.	2	n.a.	n.a.
Identifiers	18	5	9	23	13	22	20
Categories	8	11	13	11	9	14	5
Pages that link to this one	281	180	356	149	192	76	27

Table 2 shows that there are notable differences between the analysed language editions, and that each one provides only a partial view according to the language. It is noteworthy that the pages that link to a selected language edition were also different for each considered language and the majority of them originates from the same edition. The number of attracted hyperlinks by Borromini's page is also higher in the Italian version than in the English one. The German, Spanish and French Editions also contribute with a considerably high number of hyperlinks as well for Bernini as for Borromini. The results reveal special particularities according to the language editions, although they are certainly very closely associated with the interests and individual bibliographic habits of their creators. The results obtained from the online version of Encyclopaedia Britannica are summarised in Table 3.

Table 3. Analysis of Britannica's entries for both artists

www.britannica.com	Bernini	Borromini
Biography or Entry	1	1
Articles	69	17
Images	16	5
Video	0	0
Dictionary	1	0
Journals	7	4
Web's best sites	6	2
Primary Sources/E-Books	0	0
Year in Review	3	0
Additional Readings	12	3

In contrast to the English edition of Wikipedia, the content is here comparable for both artists. Bernini' entry is larger and more complete than Borromini's though. On the other hand, Britannica focuses on providing a selection of the most reliable information, for example, only

the "web' best sites". Under "additional readings" the user can also find a selection of the recommended bibliographies, all of them that are annotated. Furthermore, the identity of the creators of this information and their affiliation is provided clearly and in a transparent way.

Table 4 shows the results from WBIS Online and Oxford Reference. It is noteworthy that WBIS lists the number of biographical entries in the language archives, while all searches in Oxford Reference are sent to the Full Text. This explains the high difference between the number of items retrieved in both sources.

Table 4. Results from WBIS for both artists

WBIS Online	Bernini	Borromini
Archivio Biografico Italiano (ABI)	28	15
Deutsches Biographisches Archiv (DBA)	0	6
Biographical Archive of Christianity (BAChr)	1	0
Archives Biographiques Françaises (ABF)	1	0

Excluding the Italian Archive, Bernini appears in two additional archives (BAChr and ABF), Borromini only in one. It seems paradoxical that Bernini appears in the archive of Christianity when Borromini is not, since the latter was the most fervent and devout Christian. Borromini's entry in the German Archive is explicable due to his high popularity in Switzerland, where he was featured on the 6th series of the 100 Swiss Franc banknote, which was in circulation from 1976 until 2000[4]. On the other hand, Bernini's entry in the French Archive is based on his politically forced visit to France, where he was working for King Louis XIV, who required an architect to complete the royal palace of the Louvre (Morrissey, 2006).

All these results should be considered highly relevant and reliable. The same applies for the results collected across Festschriften (see Table 5).

Table 5. Analysis in IJBF entries for both artists

IJBF (Festschriften)			
Search for	**Bernini**	**Borromini**	**Bernini & Borromini**
Honorated	8	1	0
su=subject headings	22	10	1
Festchrift	8	1	0
ti=article title	38	10	1
ft= full text	63	13	1

All these sources show the same trend: Bernini attracts many more mentions and rubrics than Borromini.

C) Databases

The results obtained from the databases selected for this study, PQDT, JSTOR[5] and ABM (Table 6) are compared with the ones obtained in the classical bibliometric sources, WoS CC

[4] This decision at that time caused polemics in Switzerland, started by the Swiss Italian art historian Piero Bianconi. According to him, since in 17th century the territories which in 1803 became the Canton Ticino were Italian possessions of some Swiss cantons (Condominiums of the Twelve Cantons), Borromini could neither be defined Ticinese nor Swiss.

[5] Note that approximately 10% of articles on JSTOR have abstracts. To widen the search it is necessary to remove the abstract filter to search article full-text.

and Scopus (Table7), as well with the ones obtained from Google Scholar, Microsoft Academic and CrossRef via Publish or Perish (Table 8).

Table 6. Results from PQDT, JSTOR and ABM

Database	PQDT			JSTOR			ABM		
Search for	**Bernini**	**Borro mini**	**Bernini & Borromini**	**Bernini**	**Borro mini**	**Bernini & Borromini**	**Bernini**	**Borro mini**	**Bernini & Borromini**
su=subject/keyw	18	1	0	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
ti=title	18	3	0	296	51	0	57	5	0
ab= abstract	60	15	7	78	19	3	103	16	3
su OR ti OR ab	67	16	7	343	61	3	129	19	3
ft= full text	*3899*	*522*	307	*9132*	*1860*	498	453	63	29
ft not au	*3897*	*522*	307	*8874*	*1859*	498	445	63	29
ft refine exclude natural sciences	*3090*	*522*	307	*8349*	*1857*	498	445	63	29
ft only closed related topics	*687*	*339*	236	*5718*	*1507*	498	444	63	29

Table 7. Results from Web of Science Core Collection and Scopus

Web of Science Core Collection				Scopus			
Search for	**Bernini**	**Borro mini**	**Bernini & Borromini**	**Search for**	**Bernini**	**Borro mini**	**Bernini & Borromini**
keyword+	9	1	0	index keyword	2	1	**0**
author keyword	13	8	1	author keyword	26	12	1
ti=title	468	91	5	ti=title	130	28	1
ab= abstract	54	20	1	ab= abstract	99	34	3
topic	506	102	6	topic	194	52	4
cited au= docs	*544*	*77*	n.a.	secondary documents	*1487*	*376*	29
cited au= citations	*653*	*90*	n.a.				
cited au=citing docs	*n.a.*	*59*	3				
cited work=docs	*1498*	*301*	n.a.				
cited work=citations	2526	495	n.a.				
cited work= citing docs	1067	256	71				

In the case of Google Scholar and CrossRef the most appropriate search option was the search in Title, because the search in “All the words” also included publications from homonymous authors but not referring to our person. To exclude them manually would require exhaustive work that would not justify its value or relevance for this study. A maximum of 1000 results can be retrieved per search. In the case of the search for Bernini in Title (<1000), the results were downloaded in four tranches of around 400 items. The searches for Borromini and “Borromini AND Bernini” could also be performed successfully in the field “All the words”. For Borromini in “all words” 14 downloads were necessary. The corresponding results for Google Scholar from Table 8 show clearly the difference between high "precision" and relevance (Search in Title) and high "recall" but minor relevance as a simple mention in the full

text document. In Microsoft Academic the research for all the words did not include the author field and could therefore be applied more easily.

Table 8. Results from Google Scholar, Microsoft Academic and CrossRef via Publish or Perish

Google Scholar								
Search for	**Indicator**	**Bernini**	**Borromini**	**Bernini & Borromini**	**Search for**	**Bernini**	**Borromini**	**Bernini & Borromini**
Title	Papers	1649	505	37	**All words**	n.a.	10462	4448
	Citations	5398	1301	80		n.a.	96026	49092
	Cites_Paper	3.27	2.58	2.16		n.a.	9.18	11.04
	Max. Cites	311	78	19		n.a.	3454	2368
	First year	1713	1713	1957		n.a.	1702	1725
	Last year	2018	2018	2016		n.a.	2018	2018
Microsoft Academic					**CrossRef**			
All words	Papers	670	165	9	**Title**	197	45	9
	Citations	315	108	3		33	21	0
	Cites_Paper	0.47	0.65	0.33		0.17	0.47	0
	Max. Cites	21	19	3		4	8	0
	First year	1855	1967	1983		1886	1969	2014
	Last year	2018	2018	2016		2018	2018	2017

All the results show the same trend: Bernini attracts many more mentions than Borromini either in Title or in Topic or in the full text. The results confirm the lower coverage of Scopus and WoS CC, especially in comparison with JSTOR and Google Scholar. Google Scholar is the data source providing the highest scores (papers and citations) for both the search in Title and in full text. The results from Microsoft Academic are considerably lower than the ones from Google Scholar (Sven, 2017). The results from CrossRef are very similar to the ones resulting from Scopus.
All these results hint at the urgent need to include these data sources in the analyses grasping at assessing the broad impact.

D) Web tools
The results gained from YouTube and Twitter are summarised in Table 12. For YouTube they include the number of videos retrieved, the number of views as well as short statistical résumé containing maximum and minimum, mean and median, and standard deviation. For Twitter, the number of tweets, Replies, Retweets and Likes.
YouTube retrieved 473 uploaded videos dealing with Bernini and Borromini with over three Million views. The maximum of views is reported in 2012 (37 videos with 780040 views), the maximum of videos uploaded in 2017 with 68 (in 2018 already 67, see Figure 5). Comparing with the data resulting from Wikipedia and limiting to the same period (2015-2018) the views in YouTube also reach a peak in 2016[6].

Table 12. Results from Youtube and Twitter

YouTube		Twitter: 177 tweets		
# Videos	473	Replies	Retweets	Likes

[6] Note that Figure 5 is represented on a logarithmic scale.

Total Views	3607983	Total	77	583	1106
MaximumViews	254724	Maximum	4	381	387
Minimum Views	2	Minimum	0	0	0
Mean	7627.87	Mean	0.44	3.29	6.25
Median	1695	Median	0	0	1
Standard Deviation	22094.31	Standard Deviation	0.80	28.72	31.41

While the information originating from YouTube was found of high interest, the one originating from Twitter was almost reduced to visitors' likes or displays of admiration in front of their works of art. The number of replies in Twitter was extremely low, but therefore contained interesting background information[7].

Conclusions

Without doubt historical celebrities like Bernini and Borromini are a good choice for a bibliometric study in order to reveal appropriate data sources for broad impact assessment in the scholarly community and beyond the „scholarly realm“. Both artists and architects have left a rich legacy for posterity. Our study corroborates that their works have lost none of their timeliness and regency throughout the centuries and continue to be obligatory references in the world of the Arts and Humanities.

Our results also clearly show that sources normally not considered for comparable bibliometric analyses, like JSTOR, PQDT, Google Scholar, etc. in fact provide more relevant or abundant information than the usual suspects, like the Web of Science Core Collection and Scopus.
Today we are forced to respond to the manifold challenges of the digital and virtual eras, and we therefore constantly struggle to expand our data universe in order to paint a more complete picture of the broad impact assessment of individuals. It is therefore crucial to identify the most essential and appropriate data sources for each discipline, and always critically challenge their completeness, suitability and efficiency.
It is definitely insufficient to count mentions and others signals collected in blogs, social media and further tools obtained from Web 2.0. Citations and/or mentions derived from databases are still important. Particularly for subjects in Arts and Humanities it is important to broaden the scope of data sources with regard to subject specificity and coverage of other document types than research articles. Other document types, like theses or monographs should not be missing in any attempt to estimate the broad impact generated especially in the disciplines of social sciences, arts and humanities.

Moreover, our study puts in question the completeness of sources such as Wikipedia. In the reduced way how tools like Altmetric.com or PlumX already exploit this data source, we actually lose much of the new opportunities and the rich information they could provide, especially those related to different languages and cultures. The differences observed between the English and Italian Wikipedia editions for Bernini and Borromini are certainly not casual. Reducing the counts to the English version of Wikipedia cannot be regarded as best practice and potentially hampers the reliability of the assessment. Page views for each language edition in Wikipedia should at least be regarded equally or even more significant than the number of tweets or likes in social media for the assessment of the attention a subject has received on the web (see also Katz and Rokach, 2017).

[7] Like the one commenting Canaleto’s painting of the Pantheon depicting the incongruent two bell towers, which Romans dubbed l'orecchie d'asino or "ass’s ears" and wrongly attributed to Bernini

New data sources can of course be considered in a complementary way, but their significance should always be challenged and checked. According to our results YouTube can be a very rich and promising data source in addition to the world of publications. Nevertheless, there are some issues to be tackled, like especially the instability of the data and the poor syntax not allowing to perform a precise search. The significance of Twitter turned out to be very low, though.
"Publish or perish" has proven to be a valuable tool for tracking mentions, but the syntax does not yet allow conducting complex searches.
Both the classic and the new data sources show much room for improvement in order to be used more efficiently, and further studies of this kind are necessary to make more well-grounded statements.

The process of citing or mentioning is a process of equal parity, since it is one publication that cites or mentions another one, and the two are comparable. The situation is quite different in the realm of new metrics, though. Here it is a user - sometimes not even the author of anything - who views, downloads, comments or discusses a publication (Gorraiz, 2018). Therefore, there is a danger that these new metrics open the door to a radical change in the sciences and rather turn them into a marketing competition than focusing on true merits. On the other hand, this whole new internet universe has also exploded the number of indicators that we can collect quickly and easily. Being aware that nowadays a new publication appears every second, and this in turn generates an endless number of signals we face without exaggeration, a new danger: the "Tower of Babel" effect, for giving it a biblical accent (Gumpenberger et al., 2018). Curiously, the lantern of Sant' Ivo is topped with a spiral, and therefore brings to mind this very tower of babel, another ancient (if not counterintuitive) symbol of wisdom. Apparently Borromini wanted to warn us that both concepts are closely linked to each other (Morrissey, 2006).

Finally, the winner of both Baroque opponents may be Bernini (and his footsteps and shadows may be more numerous and international than Borromini's). However, it is noteworthy that the rivalry between these two geniuses has already become legendary. It is well reflected in one of the tweets analysed during this study: "Ronaldo and Messi are this generation's Bernini and Borromini".

Acknowledgments
The authors thank Jack Morrissey for his wonderful book about the rivalry that transformed Rome. This book has been the source of inspiration and the great incentive for this work. The authors are also grateful to Christian Gumpenberger for his helpful language editing.

References

Andersen, J. P. and Hammarfelt B. (2011). Price revisited: on the growth of dissertations in eight research fields. *Scientometrics,* 88(2), 371-383. doi: 10.1007/s11192-011-0408

Blunt, A. (1979). *Borromini*. Belknap Press of Harvard University Press, Cambridge MA. ISBN 0-674-07926-4.

Mormando, F. (2011) *Bernini: his life and his Rome*, Chicago, Ill. [u. a.]:Univ. of Chicago Press, 2011, ISBN 978-0-226-53852-5

Burbaum, S. (1999). *Die Rivalität zwischen Francesco Borromini und Gianlorenzo Bernini.* Oberhausen: Athena. 1999.

Giles, J. (2005). "Internet encyclopedias go head to head". *Nature*, 438 (7070): 900–901. doi:10.1038/438900a.

Gorraiz, J., Gumpenberger, C., & Wieland, M. (2011). Galton 2011 revisited: A bibliometric journey in the footprints of a universal genius. *Scientometrics*, 88(2), 627-652. 10.1007/s11192-011-0393-y Retrieved from www.scopus.com

Gorraiz, J.; Wieland, M.; Gumpenberger, C. (2017). “To be visible, or not to be, that is the question”. *International journal of social science and humanity*, 7 (7), 467-471. https://doi.org/10.18178/ijssh.2017.V7.868

Gorraiz, J., Gumpenberger, C., & Wieland, M. (2015). Galton 2012: the bibliometric journey continues. Qualitative & quantitative methods in libraries, 559-573

Gorraiz, J. (2018) A Thousand and one reflections of the publications in the mirrors`labyrinth of thenew new metrics. El professional de la información. 27 (2), 231-236. http://doi.org/10.1007/s11192-016-19991-5

Gumpenberger, C.; Glänzel, W.; Gorraiz, J. (2016). “The ecstasy and the agony of the altmetric score”. *Scientometrics*, 108, 977-982; htps://doi.org/10.1007/s11192-016-1991-5

Harzing, A. W. (2007). Publish or perish. http://www.harzing.com/pop.htm

Katz, G and Rokach, L. (2017) Wikiometrics: A Wikipedia Based Ranking System. *World Wide Web*, 6. https://doi.org/10.1007/s11280-016-0427-8

Marx W., Cardona M., and Lockwood D. J. (2011). Rutherford's scientific impact from a bibliometric perspective. *Australian Physics*, 48(3), 78-83.

Moed, H. F. (2017). *Applied evaluative informetrics*. Springer International Publishing. ISBN: 978 3 319 60521 0

Moed, H. F. (2017). From Eugene Garfield’s Citation Index to Scopus and Google Scholar. Chapter 14. In: *Applied evaluative informetrics*. Springer International Publishing. ISBN: 978 3 319 60521 0, p.193-2015

Morrissey, J. (2006) *The Genius in the Design. Bernini, Borromini, and the Rivalry That Transformed Rome*. Harper Perennial, Collins e-books. Kindle-Version. ISBN-13: 978-0060525347

Hug, S. E., Ochsner, M. and Brändle, M.P. (2017). “Citation analysis with microsoft academic.” *Scientometrics,* 111, 371-378.

Wikipedia contributors. (2018, December 19). Wikipedia. In Wikipedia, The Free Encyclopedia. Retrieved 21:04, December 20, 2018, from https://en.wikipedia.org/w/index.php?title=Wikipedia&oldid=874419259

RESEARCH BEYOND SCHOLARLY COMMUNICATION – THE BIG CHALLENGE OF SCIENTOMETRICS 2.0

Wolfgang Glänzel[1,2], Pei-Shan Chi[3]

[1]*wolfgang.glanzel@kuleuven.be*
ECOOM and Dept. MSI, KU Leuven, Naamsestraat 61, 3000 Leuven, , Belgium

[2]*glanzw@iif.hu*
Library of the Hungarian Academy of Sciences, Dept. Science Policy and Scientometrics, Budapest, Hungary

[3]*peishan.chi@kuleuven.be*
ECOOM, KU Leuven, Naamsestraat 61, 3000 Leuven, , Belgium

Abstract

In the present study we discuss the challenge of "Scientometrics 2.0" as introduced by Priem & Hemminger (2010) in the light of possible applications to research evaluation. We use the Web of Science subject category *public, environmental & occupational health* to illustrate how indicators similar to those used in traditional scientometrics can be built, and we also discuss their opportunities and limitations. The discipline under study combines life sciences and social sciences in a unique manner and provides usable metrics reflecting both scholarly and wider impact. Nonetheless, metrics reflecting social media attention like tweets, retweets and Facebook likes, shares or comments are still subject to limitations in this research discipline as well. Furthermore, Usage metrics clearly point to the manipulation proneness of this measure. Although the counterparts of important bibliometric indicators proved to work for several altmetrics too, their interpretation and application to research assessment requires proper context analysis.

Background and introduction

Scientific communication was long dominated by scholarly communication in basic research in the sciences. Scholarly communication in fundamental research, in the so-called "hard sciences", took and partially still takes place in journal literature. A simple look into typical documents in these research areas reveals this kind of communication patterns: Most references point to other journal articles and scientific literature in periodicals thus covers most sources and targets of scholarly communication. The mission of the Scientometrics 1.0 version, which came up in the 1960s-1970s of the last century, was to model and measure documented scholarly communication in basic science and impact on scientific communities. The identification of both actors and users of scientific information was easy, as those could be found within the scientific communities. The even sometimes disputed use of citation measures for evaluative purposes was rather clearly defined as citations marked the information use in the process of knowledge production and dissemination with well-defined rules.

When scientometrics opened towards new data sources (including conference proceedings and books) and broadened towards the measurement of research performance in other fields than basic research, it became apparent that the above-mentioned framework proved too narrow for those fields. Researchers pointed to the fact that, e.g., in medical and applied sciences, a large share of information targets is found outside the research community and citations are therefore yet partial measures of impact and use of information. In the social sciences and humanities (SSH), a large share of both sources and targets are located outside the research community, thus citations based on periodicals can only be considered an insufficient measure of impact and use of information. Scientometricians attempted to catch up with this challenge and to keep pace with the new developments in research evaluation by broadening the scope and improving their methods.

Yet, the new millennium came up with new challenges to be met by scientometricians. These challenges partially result from the new demands through policy and society needs, new

movements, like 'open science', are also caused by the new electronic communication forms becoming prevalent in scientific communication as well. These intra-scientific, societal, policy-driven and technical demands lead to the evolution of a new concept called "Scientometrics 2.0" (Priem & Hemminger, 2010). Priem and Hemminger considered *open science*, *social media metrics* and *alternative metrics* groundwork and components for this new concept. They also compiled a list of possible sources for its implementation. Yet, as so often reality is quickly running past visions and nowadays a plethora of measures and metrics are in use, sometimes in a rather uncritical manner and even repeating or imitating typical errors of the early Scientometrics 1.0 (Gumpenberger et al., 2016).

Previous results and research questions

Like in our previous studies special focus is laid on the comparison with traditional, mainly publication and citation based indicators. In addition, we will show that several advanced methods and indicators developed for traditional 'productivity' and citation analysis are still fit for the new environment. From the conceptual-methodological viewpoint, five categories, namely, *Usage, Captures, Mentions, Social Media* and *Citations* are distinguished according to PlumX (PlumX, 2018). These five categories still form just a minor part of what can be covered by *broader* or *societal impact* (cf., Lewison, 2004, 2008). On the basis of our own research (e.g., Chi & Glänzel, 2018, Chi et al., 2018), we will focus on three of the five groups of metrics, particularly, on Captures, Citations and Usage.

For the present study we use publication, citation data and usage statistics from Clarivate Analytics' Web of Science Core Collection (WoS) in conjunction with altmetrics data from Plum Analytics. In our previous studies, we analysed selected fields from the sciences and social sciences to uncover specific patterns of impact, information ageing. Thus the results could readily be compared with those of traditional scientometric studies. We could also show that (full-text) download processes generally mirror the characteristics of citation processes but not always to the same extent and mostly with a certain field specific "translation coefficient" (cf. Glänzel & Heeffer, 2014). This implies that one citation roughly corresponds to a certain number of downloads, which amounted to about 100 in our Elsevier sample of 80,000 journal documents put online in 2008 and followed up for downloads and citations with a five year window. The citation process mirrors the increments of downloads, however with a certain 'phase shift' in accordance to our expectations. The correlation between the impact and the usage measure proved very strong, which partially confirmed results of earlier studies by others (e.g., Moed, 2005; Brody et al., 2006; Thelwall, 2012). Further studies by Chi and Glänzel and most recently, by Chi et al. (2018) could confirm and deepen these results. We could also show that traditional concepts and methods can be integrated into the new metrics. We defined a *Journals Usage Index* (Chi & Glänzel, 2018) in analogy to the Garfield Impact Factor as well as the idea of relative citation indicators, and the Characteristic Scales and Scores proved to work for new metrics as well.

In our previous studies we could already make some specific observations. The most important one concerns the difference between the patterns in basic research and the social sciences and humanities. In terms of WoS usage statistics of journal articles, social sciences displayed disproportionately higher "usage" than citation impact (Chi & Glänzel, 2018, 2019). This did not strike us because citations to periodicals play a less pronounced part than in the sciences (Chi & Glänzel, 2019). All the more, we found it interesting that the usage of authored and edited books did not reflect the same patterns (Chi & Glänzel, 2019). Figure 1 gives the correlation between the mean usage rate (MUR) and the mean citation rate (MCR) of two document types of book publications, authored and edited books as reflected by the 2013 volume of Clarivate Analytics *Book Citation Index* (BKCI).

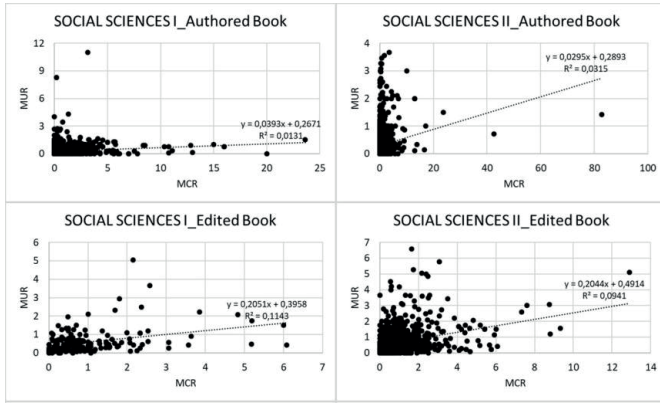


Figure 1. Scatter plots of MUR vs. MCR for two document types in two major fields

In the present study, we will therefore further elaborate our methodology for the application and systematic analysis of an interdisciplinary field connecting both the life sciences and the social sciences. In particular, we have chosen the WoS subject category "*public, environmental & occupational health*". Furthermore, this discipline has already attracted our interest in terms of its growth and its emerging topics (Glänzel & Thijs, 2011, 2012).

In the light of previous results, we will attempt to answer mainly the following research questions.

1. In our previous studies we have found different extents of correlation between scholarly impact and Usage/Captures metrics with regard to disciplines and publication types. Will we find similar patterns as has been found in the (life) sciences?
2. Will the altmetric indicators distinctly exceed the scholarly impact with a factor for "translating" impact to Usage and Captures amounts?
3. Can we observe specific national patterns and can we find journals with significant deviation of their altmetrics from their traditional bibliometric characteristics?

In addition to the counts, shares and mean values, we will also use Characteristic Scores and Scales (CSS; Glänzel & Schubert, 1988; Chi & Glänzel, 2018) to analyse distributional aspects of the metrics and to identify extreme values and outliers.

Data sources and data processing

For this study we have selected the WoS subject category *public, environmental & occupational health* ("public health" in short) in the Social Sciences Citation Index (SSCI) to emphasise the social sciences interaction in the selected field. All documents published in 2015 and indexed as articles and reviews in the SSCI have been downloaded from Clarivate Analytics' WoS database. All papers have been assigned to countries on the basis of the authors' corporate address and to the journals in which the papers have been published. Citations and usage have been counted till October 2018, that is, based on three years on an average. For matching the WoS data with altmetrics, the document DOIs and the PubMed IDs have been used. In total, 18,729 out of 18,824 articles and reviews, that is, 99.5% of all retrieved documents proved to have a valid DOI or PMID.

In a second step Usage, Captures and Citation metrics have been downloaded from Plum Analytics (https://plumanalytics.com/) in October 2018 for each individual document using its DOI or PubMed ID. Possible errors in the identifiers have been corrected manually. Data downloaded from WoS and PlumX have been carefully cleaned and processed to bibliometric indicators.

Methodological considerations

PlumX organises metrics into five categories where the categories represent different levels of information use. "Usage" stands for the lowest level. This comprises, for instance, clicks, downloads or views and could rather be considered measures of the intention to use something than their actual usage (Gorraiz et al., 2014). "Captures" expresses somewhat more as it indicates repeated usage, for instance, as bookmarks, favourites, readers or watchers. The category "Citation" representing the highest level can be considered an extension of the concept of citations within the framework of traditional bibliometrics as this category reaches out beyond the framework of scholarly communication. In a previous study, we have not used the categories "Mentions" and "Social media", which partially require full-text and because of the document unavailability to a broader community and the "zero inflated" distributions resulting from it; therefore, we decided to omit these two indicators (cf. Chi, Gorraiz & Glänzel, 2018). In particular, the situation for "Mentions" we have encountered in public health did not essentially differ from that in chemistry and mathematics and although the patterns of "Social media" are much more favourable than in the three subjects in the sciences, we cannot consider this metrics as a candidate for application to all disciplines. However, the distinct perspective of "Social media" representing the social impact deserves a further investigation. We include this metrics in this study to broaden the discovery based on alternative metrics despite its very low percentage of data availability.

On the basis of the metrics selection we obtained a set of twelve metrics, Usage, Captures, Social Media and individual metrics Scopus and CrossRef citations, EBSCO Abstract-views, EBSCO full text views, EBSCO link-outs, Tweeter and Facebook from PlumX and Usage Count and Times Cited from the Web of Science. We applied our standard statistics, zero frequencies, mean values and Characteristic Scores and Scales to this set of metrics and we have broken down the data to the country level and to individual journals. In addition, we have conducted a regression analysis in order to detect possible correlation between these metrics and to find the "translation" coefficient, provided the correlation is strong enough.

Results

Regression analysis

Similarly to our previous studies, first we have applied a linear regression analysis to study the relationship between the selected altmetric indicators and the traditional measures of scholarly impact. Instead of the Pearson correlation we have again applied Spearman correlation because of the skewed distributions underlying all metrics. The results of the regression analysis are presented in Figure 2. Similar to the results we found in Chi et al. (2018), the three citation indicators correlate with each other strongly, especially between Scopus and WoS. Captures and WoS Usage have moderate correlations with citation indicators. PlumX Social Media has moderate to weak correlation with all the other metrics.

PlumX usage shows the most distinct patterns from other indicators. It only correlates strongly with PlumX captures indicator and has weak correlations with all the citation and Social Media indicators. The three EBSCO usage indicators correlate with PlumX captures and WoS Usage at moderate to strong level, and have weaker correlation with citation and social media indicators. Among the three usage indicators, EBSCO full text views has the most distinct patterns from others. It is extremely weakly correlated with other indicators except for EBSCO abstract views. Its weak correlation with other metrics is even weaker than that of Facebook metrics with others which has, otherwise, also the highest share of zero-frequencies.

Another interesting finding is,that the social media metrics is stronger correlated with citations than Usage and Capture. This result is different from the previous studies of the correlation between citation and altmetric indicators (e.g., Costas, Zahedi & Wouters, 2015; Zahedi, Costas & Wouters, 2014). The message conveyed by Figure 2 substantiates the different dimensions of

usage, captures, social media and citation measurements and their relative degrees to the scholarly contribution.

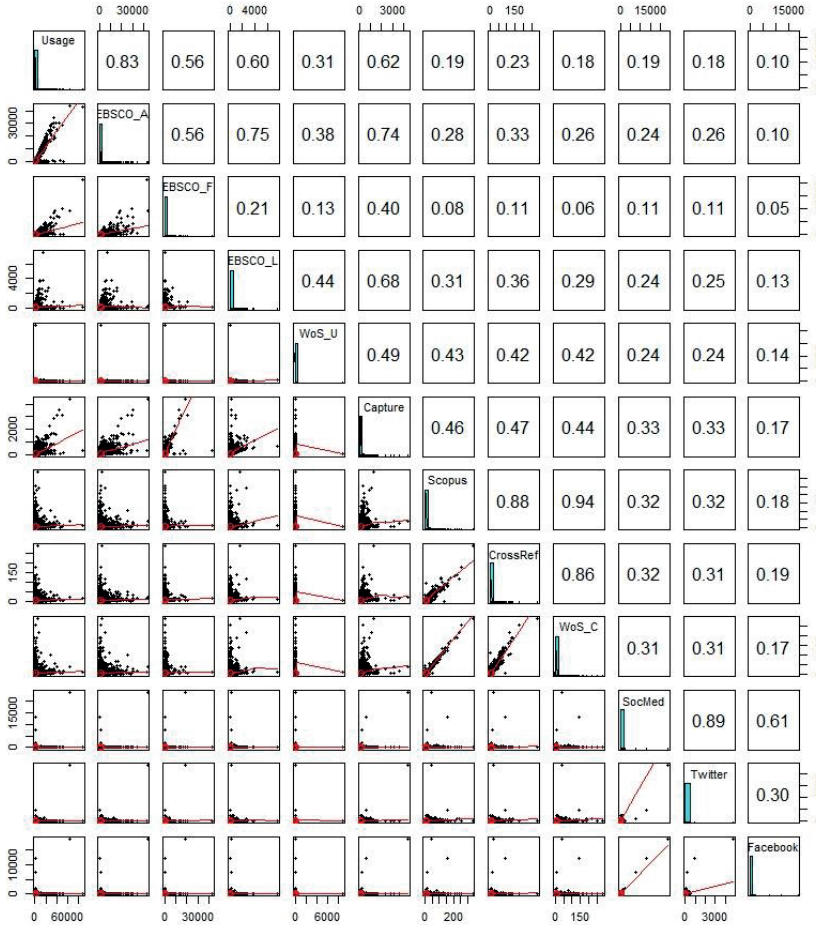


Figure 2. Spearman correlation between twelve metrics for the 18,729 documents in public health

EBSCO_A: EBSCO Abstract-views; EBSCO_F: EBSCO full text views; EBSCO_L: EBSCO link-outs; WOS_U: WoS usage; WOS_C: WoS ciations; SocMed: Social Media

Table 1 presents the basic indicators on the complete data set. Let X denote the random variable represented by the sample. Putting $b_0 := 0$ as the very first characteristic score, we then obtain the subsequent scores as $b_k := E(X|X \geq b_{k\text{-}1})$ for all non-negative integer values $k = 1, 2, \ldots$. These intervals $[b_{k-1}, b_k)$ between two adjoining scores define the performance classes from "poor" (Class 1) to "outstanding" (Class 4). The share of uncitedness is in line with the expectations. About one tenth of the documents remains uncited in WoS and Scopus in the period of roughly three years. Only the share of uncited documents according to CrossRef is distinctly larger. The zero-usage share on WoS platform is much lower than citations. The share of zero-captures is almost minute and even much lower than the shares of PlumX usage and social media and the other aggregated categories containing several individual metrics. The three EBSCO usage metrics keep the same zero-usage share as citations, except for the full text views,

which may be the consequence of partial unavailability of the full texts of the underlying documents. Social media metrics, unsurprisingly, have the highest zero-frequencies.
The mean values reveal even more interesting patterns. Note that the mean value coincides with the first CSS score b_1. While all citation means are of the same order, Usage and Social Media reflect completely different patterns. The considerable difference between the mean value of PlumX Usage and those of other EBSCO usage metrics reflects another main component of usage metric apart from EBSCO: SciELO. Even though we do not report the SciELO metrics due to their very low percentages of data availability, which is probably caused by the regional coverage, the extremely high usage of some articles in SciELO lifted up the average of general usage metrics on PlumX. The same could be observed in the case of Social Media metrics as well, however to a much lesser extent.

Table 1. Bibliometric indicators calculated on the twelve metrics of the 18,729 documents in public health

	Usage	*Abstract Views*	*Full tekst Views*	*Link-outs*	*WoS Usage*	*Cap-tures*	*Scopus Cites*	*Cross-Ref*	*WoS Cites*	*Social Media*	*Twit-ter*	*Face-book*
f_0	6.8%	7.3%	49.4%	10.1%	2.7%	1.4%	10.1%	19.1%	10.8%	45.3%	51.0%	78.7%
b_1	1001.94	601.66	108.51	64.87	11.98	80.74	7.07	4.67	6.31	10.00	3.80	6.20
b_2	3662.68	2322.30	533.34	200.60	23.98	172.27	16.48	10.63	14.56	63.92	16.30	82.23
b_3	7985.20	6112.24	1499.92	425.23	42.10	302.20	30.00	19.29	26.42	293.59	51.88	465.69
CSS1	78.3%	79.5%	81.7%	74.4%	65.0%	66.8%	69.9%	66.3%	69.0%	86.3%	79.5%	92.8%
CSS2	15.4%	15.4%	14.0%	18.4%	24.5%	23.2%	21.3%	23.5%	21.7%	11.6%	16.4%	6.2%
CSS3&4	6.3%	5.0%	4.3%	7.2%	10.6%	10.0%	8.9%	10.3%	9.3%	3.9%	5.3%	2.9%

The citation distributions have their specifically skew shape. By contrast, the distribution of captures is very flat with typical probabilities around 1% each for $0 \le k \le 50$ captures. The other half is distributed less evenly between 51 and the maximum of 4416 captures. The two usage metrics show, however, large discrepancies. The Usage distribution is extremely flat with maximum usage frequency at $k = 0$ with more than 200 extreme values, each being larger than 10,000. By contrast, the WoS usage has a more "regular" flat-tailed distribution with the mode of relative frequency of 7% at $k = 6$. We have found only one single outlier ($k = 8791$) here.
The properties of these distributions are also reflected by the CSS scores ($b_1, \ldots, b_3$) in Table 1. The three EBSCO usage metrics follow by and large the patterns of Usage. The PlumX Usage and Captures scores are, in fact, of a one or two orders of magnitude higher than the citation scores and the WoS usage counts. The three citation distributions, indeed, substantiate strong relatedness. The negative binomial distribution-model described shape and characteristics of the empirical distributions quite well although the fit to the individual frequencies is, because of the long stretched distribution, not perfect. This relatedness is also reflected by the parameters N and P of the negative binomial distribution (see Table 2) with

$$p_k = \binom{N+k-1}{k}\left(\frac{1}{P+1}\right)^N\left(\frac{P}{P+1}\right)^k, k = 0, 1, 2, \ldots .$$

N values greater than 1 indicate that the modus of the distribution might be at $k > 0$, which is the case for the WoS and Scopus citations as well as for the WoS Usage counts. The parameters of these distributions along with those of the CrossRef citations are in the same range. The large P value of the Captures distributions reflects its flatness and the extremely high P value of the PlumX Usage metrics substantiates that the negative binomial model here actually fails. Although this effect seems to be caused by the regional effect expressed by the Scielo data, the polarised distribution pattern also holds for the EBSCO Views, albeit to a much lesser extent. Only Link-outs have an almost acceptable distribution, which is quite "close" to the case of

Capture metrics. The (*N, P*) parameter-value pair of the Social Media metrics reflect interesting details. While the Twitter metrics still follow a regular, however very skewed distribution, the distribution of Facebook metrics can already be considered degenerate.
It is interesting to observe that the CSS classes nevertheless are in the same range for all indicators; only the PlumX Usage metrics shows a distinct deviation from the other ones. However, the range of all classes indicates that the CSS method is applicable in all cases. We will use this method later on, in the context of journal and country statistics.

Table 2. Parameters of the negative binomial distributions fitted to the twelve metrics for public health

Metric	*N*	*P*
Usage	0.34	2978.55
EBSCO Abstract Views	0.35	1712.07
EBSCO Full text Views	0.10	1074.72
EBSCO Link-outs	0.46	139.72
Capture	0.96	84.03
Scopus	1.18	5.09
CrossRef	0.92	6.00
WoS Citations	1.22	5.18
WoS Usage	1.74	6.88
Social Media	0.20	49.46
Twitter	0.24	15.98
Facebook	0.05	125.40

Among the 18,792 documents, some articles with *extreme values* in one indicator are not necessarily to have high values in other indicators. Table 3 lists the most extreme values of each indicators, and shows that those articles with high usage values come more often with high captures values while extremely highly cited articles keep their dominant positions among all the three citation sources. The most used articles on the WoS platform is a special case that only initiates high usage within the database but does not show high influence anywhere else (see Document #9). Pars pro toto, we will have a look at #1 with outstanding PlumX usage counts and otherwise low EBSCO usage and citation rates. #2 has very high PlumX Usage, Capture and Social Media metrics. Documents #6 – #8 have attracted above-average PlumX usage (CSS Class 3) and can be considered outstandingly cited (CSS Class 4).
The language of document #1 is Portuguese although there is an English version as well. This paper by Brazilian authors is entitled "The field of Collective Health in Brazil: definitions and debates on its constitution" and its topic and its strong regional/local focus might explain the enormous attraction in terms of usage, on the one hand, and the discrepancy between PlumX usage and EBSCO usage metrics and citation impact on the other hand. The high PlumX usage are contributed by the usage on SciELO platform. Document #2 entitled "Mental Illness, Mass Shootings, and the Politics of American Firearms" was cooperated by USA authors and drew a lot of attentions on social media and content provider platforms. This may be because of its topic coordinating public concerns and shows its societal influence resulting a vigorous discussion in society than in academia.
#6 is a review article on "The Prescription Opioid and Heroin Crisis: A Public Health Approach to an Epidemic of Addiction", which already presages the general interest and citation attractivity. The paper is published by five US institutions and, being a review it is also expected to exhibit higher citation rates than research articles. The research paper entitled "Purposeful Sampling for Qualitative Data Collection and Analysis in Mixed Method Implementation

Research" (document #7) is the result of a collaboration of six US institutions and funded by the National Institute of Mental Health. Although it has a different document type, it shows very similar patterns in terms of Usage, Captures and citations as the previous one.
Finally, document #3 is another co-publication of US institutes with highest PlumX usage, which is mainly contributed by EBSCO usage metrics, we have found in our data set. Captures and Social Media are high but not extreme. By contrast, the citation rates it has received are rather moderate (cf. Table 1). The same applies to the WoS usage count. This document is a research article on "Social Science Collaboration with Environmental Health". The importance and strong social attraction of this topic speaks for itself. These five examples may just illustrate the effect of the thematic peculiarities of research in this specialty. In the following subsections we exclude WoS Usage count as we have already reviewed this metrics and its relationship with citation measures in previous studies (Chi & Glänzel, 2018, 2019; Chi et al., 2018).

Table 3. Extreme values of the twelve metrics of the 18,792 documents in public health

#	*DOI*	*Usage*	*EBSCO_A*	*EBSCO_F*	*EBSCO_L*	*WoS U*	*Captures*	*Scopus*	*Cross Ref*	*WoS_C*	*Social Media*	*Twitter*	*Facebook*
1	10.1590/S0104-12902015S01018	46897	361	172	12	6	72	1	2	6	0	0	0
2	10.2105/AJPH.2014.302242	63131	43780	18295	263	75	4416	41	21	38	23586	4746	18840
3	10.1289/ehp.1409283	86660	43208	43328	32	12	269	14	12	11	83	5	78
4	10.2105/AJPH.2014.302393	44066	30639	13175	252	53	3597	27	12	18	2	2	0
5	10.1037/a0039045	14518	6762	173	7568	108	622	14	9	13	0	0	0
6	10.1146/annurev-publhealth-031914-122957	7694	4818	4	2528	101	1093	240	114	218	149	137	12
7	10.1007/s10488-013-0528-y	4467	2249	26	1644	81	1295	335	287	286	6	3	3
8	10.1093/ntr/ntu191	813	538	178	87	41	261	256	174	252	2	2	0
9	10.1111/jrh.12095	1217	1180	3	34	8791	55	7	4	9	0	0	0
10	10.1111/jrh.12078	904	851	2	51	332	143	2	1	2	0	0	0
11	10.1016/s2214-109x(15)70002-1	742	236	0	104	77	542	157	90	141	13106	881	12225

Journal analysis

This subsection deals with a concise journal analysis. We have selected journals with at least 250 papers of document type *article* or *review* in 2015. Eleven journals met this criterion. In order to facilitate tabulating, we use the following official acronyms in the following: AAP (Accident Analysis and Prevention), AB (Aids and Behavior), AJPH (American Journal of Public Health), APJPH (Asia-Pacific Journal of Public Health), BMCPH (BMC Public Health), CSC (Ciencia & Saude Coletiva), HE (Health Expecta-tions), IJERPH (International Journal of Environmental Research and Public Health), MCHJ (Maternal and Child Health Journal), QLR (Quality of Life Research) and SSM stands for Social Science & Medicine. The list comprises journals ranging from the highest through the lowest impact quartile. The similarity of citation patterns, including CrossRef, is striking. *Social Science & Medicine* and AJPH reflect the most advantageous situation in terms of citations (about 20% highly and about 50% or less poorly cited papers). Nevertheless, the three EBSCO usage metrics completely contradict these patterns (see Table 4). While eighty percent of AJPH papers are "highly used" according to EBSCO abstract views, about eighty percent of *Accident Analysis and Prevention* are "poorly used" and only less than 2% of its papers can be considered "highly used".
Another striking observation is the converse distribution between PlumX Usage and the three EBSCO usage metrics over the lowest (Class 1) and highly used (Class 3&4) papers published in *Ciencia & Saude Coletiva.*

Table 4. Percentage of lowest and highest CSS-class documents of the eleven largest journals in public health (2015)

Journal	*N*	U_l in %	U_h in %	A_l in %	A_h in %	F_l in %	F_h in %	L_l in %	L_h in %	P_l in %	P_h in %	S_l in %	S_h in %	R_l in %	R_h in %	C_l in %	C_h in %	M_l in %	M_h in %
AAP	305	91.8	1.6	79.7	1.6	100.0	0.0	70.5	5.9	63.0	3.3	45.2	17.4	44.6	16.1	45.6	17.0	83.3	1.0
AB	251	100.0	0.0	100.0	0.0	99.6	0.0	91.6	0.0	64.1	4.4	52.6	15.1	33.9	26.3	52.6	17.1	77.3	3.2
AJPH	512	10.4	74.4	7.8	82.0	8.2	55.3	90.4	1.2	21.3	46.7	50.8	20.9	48.4	19.9	50.0	21.5	72.5	3.5
APJPH	355	100.0	0.0	100.0	0.0	100.0	0.0	96.3	0.0	97.2	0.0	84.5	1.1	81.4	2.0	90.7	1.1	98.9	0.0
BMCPH	830	63.1	4.8	62.2	4.5	18.6	22.8	88.9	1.6	38.4	25.1	65.5	9.3	65.3	8.8	62.8	10.0	81.1	1.9
CSC	357	0.0	54.1	99.7	0.0	89.4	0.3	96.6	0.8	96.1	0.3	94.4	0.8	95.2	0.6	93.6	0.8	95.2	1.7
HE	269	89.2	0.0	86.2	0.0	66.5	1.5	62.8	7.8	55.8	10.0	64.3	10.4	61.7	10.8	70.6	8.9	84.0	0.7
IJERPH	343	99.4	0.3	99.4	0.3	99.1	0.3	96.2	1.2	90.4	1.7	66.2	8.7	46.9	17.5	63.6	8.5	81.9	4.7
MCHJ	295	70.2	1.4	54.9	2.0	52.2	5.8	63.7	7.5	35.6	21.7	68.8	5.4	54.9	14.6	71.5	5.8	83.4	2.7
QLR	290	76.2	0.0	57.6	0.7	71.7	1.0	89.0	0.3	73.1	3.4	65.5	9.7	43.1	19.3	63.4	11.0	90.7	0.7
SSM	626	78.4	1.6	67.4	2.1	100.0	0.0	33.5	24.1	36.3	20.0	47.3	20.1	44.1	19.8	45.4	20.9	62.1	8.3

Legend: N = number of papers, U = PlumX Usage, A= EBSCO Abtract views, F: EBSCO Full text views, L: EBSCO Link-outs, P=PlumX Captures, S=Scopus citations, R=CrossRef citations, C=WoS citations, M= Plumx Social Media. Index *l* denotes the lowest CSS class 1, index *h* the highest two classes 3&4 combined.

Table 5. Metrics means and percentage of lowest and highest CSS-class documents of the twenty countries with largest publication output in public health (2015)

Country	*N*	U_{b1}	U_l in %	U_h in %	A_{b1}	A_l in %	A_h in %	P_{b1}	P_l in %	P_h in %	S_{b1}	S_l in %	S_h in %	C_{b1}	C_l in %	C_h in %	M_{b1}	M_l in %	M_h in %
USA	8524	1127.19	77.3	6.8	885.96	73.2	8.3	98.46	59.0	13.3	8.28	64.7	11.0	7.44	63.5	11.7	11.92	85.8	2.3
GBR	1989	761,79	82.5	3.5	546.36	78.4	4.3	91.96	59.4	12.2	9.75	58.7	15.2	8.63	56.4	15.5	15.04	73.4	3.9
AUS	1639	823.65	79.5	3.9	554.13	76.8	4.0	88.96	61.0	12.4	8.04	63.7	10.8	7.18	63.1	11.3	8.06	84.1	2.3
CAN	1399	805.16	81.0	3.8	594.31	77.9	4.6	87.29	60.3	11.2	7.96	65.5	11.0	7.11	64.2	11.4	9.19	82.4	2.6
BRA	991	3837.03	22.9	39.7	139.56	96.0	0.7	38.23	90.9	2.7	4.38	87.0	3.5	3.98	86.4	3.4	18.89	91.6	2.3
NLD	732	579.73	84.7	1.1	423.57	80.3	1.8	77.86	65.3	8.6	8.88	60.2	13.1	7.89	59.6	13.5	9.63	83.6	2.0
CHN	706	418.37	90.1	1.3	311.75	86.4	2.4	52.92	81.3	5.1	7.19	66.1	8.8	6.28	67.0	9.1	3.14	95.2	1.1
DEU	624	498.14	88.6	1.6	337.17	83.3	0.0	58.05	76.8	6.3	7.21	71.8	9.3	6.35	71.3	11.1	4.43	88.3	1.0
SWE	572	539.73	88.8	2.1	384.14	86.2	2.4	72.73	70.6	6.1	7.67	68.5	9.1	6.71	68.5	9.1	8.74	83.4	2.1
ESP	516	933.23	72.7	4.1	322.90	87.8	1.7	61.16	77.3	6.8	6.47	73.3	7.9	5.73	74.2	8.3	10.39	83.1	2.5
ZAF	391	558.94	84.9	1.5	423.66	81.6	3.1	97.39	56.5	11.0	8.74	61.9	9.5	7.91	61.9	12.3	6.58	81.1	1.3
ITA	343	338.10	92.7	0.9	258.43	89.5	1.2	52.84	79.3	4.7	8.41	61.8	14.9	7.35	63.6	15.2	11.18	86.3	3.5
NOR	322	658.33	85.1	2.8	451.78	83.5	2.8	79.12	69.3	11.5	8.79	64.3	12.1	7.77	63.7	12.7	12.07	81.1	4.0
DNK	315	401.35	91.4	0.6	291.20	87.0	0.6	69.16	70.8	7.0	7.65	65.7	10.5	6.88	62.5	10.8	9.62	82.5	2.9
FRA	296	537.41	88.5	1.7	324.30	84.8	1.0	63.93	72.6	6.8	9.91	58.1	18.9	8.87	57.1	19.6	12.02	82.8	3.4
CHE	296	621.68	86.8	3.0	440.95	83.4	3.7	92.57	62.5	12.5	11.99	52.4	18.2	10.90	50.3	20.3	12.62	79.1	3.4
JPN	287	368.19	92.0	1.4	257.97	91.6	1.0	46.42	86.1	2.4	5.74	76.7	5.9	5.01	76.3	7.7	7.21	88.2	3.8
IND	261	468.15	88.5	1.5	367.25	85.1	2.3	68.50	71.3	6.5	7.46	72.8	9.2	6.26	75.1	9.6	8.18	86.6	2.7
IRN	253	181.74	96.0	0.4	128.01	93.7	0.4	30.35	89.7	1.6	2.46	90.1	2.4	3.74	85.0	4.0	5.06	97.2	0.8
BEL	251	582.36	84.1	1.6	413.61	81.7	1.6	93.69	60.6	13.9	10.17	59.0	15.1	8.78	60.2	15.1	6.66	79.7	0.4
World	18729	1002.05	78.3	6.3	611.66	79.5	5.0	80.75	66.7	10.0	7.07	69.9	8.9	6.32	69.0	9.3	10.00	86.3	3.9

Legend: N = number of papers, U = PlumX Usage, A=EBSCO Abstract views, P=PlumX Captures, S=Scopus citations, C=WoS citations, M= Plumx Social Media. Index *l* denotes the lowest CSS class 1, index *h* the highest two classes 3&4 combined. Index b_1 denotes the mean value of the corresponding metrics.

CSC journal has zero percent of papers with PlumX Usage lower than the average, which means locating in Class 1, but owns almost zero percent of papers with EBSCO usage metrics in highly used group Class 3 &4. This indicates the high SciELO usage counts of papers in this Brazilian journal.

The CSS classes for Capture reflect situations somewhere "in between" those of citation and usage. Social Media keeps its pattern of high shares of Class 1. Capture metrics of *Social Science & Medicine* are rather following its citation distribution, while those of IJERPH are closer to its EBSCO usage patterns. Since the number of papers in the selected journals is large enough (250 or more), the lack of clear correlation patterns can be considered significant without any particular test. Furthermore, these indefinite patterns intersect the individual journals as well. For instance, 32 out of the 512 AJPH papers (i.e., 6.25%) can be found in Capture CSS-Class 4 and in WoS Citation class 1. The same applies to 25 (i.e., 3.0%) papers published in *BMC Public Health*. Nevertheless, 19 (i.e., 3.7%) of the AJPH and 14 (1.7%) of the *BMC Public Health* papers are in the highest class according to both, Captures and WoS citations.

To conclude, Usage and Capture metrics supplement traditional journals impact measures but their interpretation is rather difficult while Social Media metrics keeps its own highly skewed pattern. Those cases, journals and individual publications, where metrics reflect contradicting situations, further, preferably context related analysis would be necessary for correct interpretation.

Country analysis

The last subsection is devoted to a comparative analysis of countries. We have selected those countries that have (co-)authored at least 250 papers of document type *article* or *review* in 2015. Twenty countries met this condition. For data presentation, we will use their three-literal ISO codes. For the comparison we have used only PlumX Usage, Capture, Social Media metrics, EBSCO abstract views and Scopus and WoS citation rates. In addition to the CSS classes, we have added the mean values of the metrics, which actually coincide with the corresponding CSS b_1 scores.

The indicators are given in Table 5. The citation indicators by and large reflect a well-known situation. Research in the US and several countries in West- and North-Europe exhibit high citation impact. This is reflected by both Scopus and WoS citations and includes mean citation rates as well as citation distribution over CSS classes. In particular, Switzerland, France and the UK show the most favourable patterns in terms of citation impact. Brazil and Iran form the low-end of the selection. This is contrasted by the PlumX usage indicators.

Table 6. Spearman correlations between PlumX usage and EBSCO abstract views for the five countries with largest publication output in public health (2015)

Country	ρ
USA	0.9790391^{***}
GBR	0.976407^{***}
AUS	0.9609342^{***}
CAN	0.7345188^{***}
BRA	0.007268

$^{***}p < 0.001$

Except for the US, which are slightly above the world standard (see Table 5), Brazil is the only country in the selection with usage-metric values that are considerably above the expectation. The mean value is almost four times the expectation and the share of highly used papers exceeds that of poorly used papers. However, Brazil has very low mean value of EBSCO abstract views.

The spearman correlation between PlumX usage and EBSCO abstract views for Brazil is not significant ($\rho = 0.007268$, $p > 0.05$; see Table 6). The extreme high PlumX usage value was confirmed by an additional check contributed from SciELO usage counts. The effect of SciELO usage is not that distinct in any other countries. The usage of the remaining 18 countries falls distinctly short of the expectations with relatively correlated abstract view usage. For example, Table 6 shows that the top 5 countries except for Brazil all have strong spearman correlations between PlumX usage and EBSCO abstract views.

For Brazil, we find a clear contradiction between (PlumX) Usage and EBSCO usage/Citation metrics. Without further analysis of the background and motivation of usage and its user community, this metrics does not convey any clear message. Capture, by contrast, provides a more differentiated and less polarised picture. Deviations from the reference standard are less extreme and more in line with what one would expect from an impact measure. This measure seems indeed to provide added value to the scholarly impact. The effect of the outliers in altmetrics is once more expressively shown by the comparison of Social Media metrics of Brazil and the USA (cf. Table 5). Both countries have the same percentage high Social Media mentions (2.3%) while Brazil has a distinctly higher share of low mentions (91.6% – vs. 85.8% for the USA). Therefore, one would expect a mean value of this metrics of Brazil much lower than in the case of the USA. By contrast, the opposite case can be observed. With b_1=18.9 Brazil clearly "outperforms" the USA (b_1=11.9). However, this result is the effect of one single outlier: Brazil has one document mentioned, shared or commented on more than 13,000 Facebook pages. Although the USA has also one extreme outlier (>23,000), the effect of their document is absorbed by their large publication output, which is of one order of magnitude larger than that of Brazil. In the CSS model both documents are just one item in the highest class, where their actual numerical value does not have any further effect. This example may illustrate that mean-value based altmetric indicators should be used with the utmost caution, most notably in the case of smaller publication sets.

Conclusions

The example of category *public, environmental & occupational health* has provided interesting and assumingly typical insight in the properties of altmetrics. Above all, these properties determine the opportunities and limitation for their possible application in an evaluative context. The PlumX Usage metrics provide – at least in the subject under study – usable numerical information on abstract and full-text view as well as on EBSCO link outs. Other forms of usage were less or not significant. Above all, SciELO seems to be responsible for the outstanding usage counts of the Brazilian papers in public health. The biases, the extremely flat distribution and the experienced low robustness of this metrics make it less appropriate for application in research assessment. The more robust WoS usage count lacks clear interpretation and requires access and use of the WoS database. This metrics might be an interesting companion to the WoS citation data as it leaves the scope of scholarly communication. All citation metrics proved to correlate, most notably WoS and Scopus but these are restricted to scholarly communication.

Captures and social media may have the potential to provide additional information to citation impact. The two important components of Capture were Mendeley and Exports/Save counts. CiteUlike was, however, not significant. The usefulness of Mendeley readership as early impact indicator has recently been shown by Thelwall (2018), but he also pointed to limitations for their use in research evaluation (Thelwall, 2017a,b). Tweeter is the most influential component of PlumX social media metrics, while Facebook is not that common to be used to disseminate research for most public health publications and Google + is never used. The distribution of social media metrics is very skewed with zero frequencies close to 50% in public health. This results in severe limitations for the general applicability of indicators based on these metrics.

In the present study we primary explored the possibility to compare the social media metrics to other altmetric and bibliometric indicators.

After our short digression to the world of altmetric indicators, as they represent the state-of-the-art, we can conclude that the indicators, in their present designed and availability, do not provide any comprehensive solution nor alternative to the well-elaborated and consistent system of scientometric tools, apart from those well-known conceptual and methodological limitations. Most strikingly, in this study, just like in our previous papers (e.g., Chi et al., 2018; Chi & Glänzel, 2019), we have found some lack of consistency in these measures. Adding, removing or just changing repositories or databases may result in dramatic changes and may turn local or regional effects into global phenomena. The database SciELO may just serve as an example for this effect. Just counting downloads, mentions, likes, tweeds and other social-media related measures without knowing the real purpose behind these actions certainly cannot provide unequivocally interpretable (quantitative) evidence. Once again we have to point to the insightful and profound article by Abraham Bookstein (1997) on the demons to measurement in social sciences. In his study, he characterised, in the context of informetrics distributions, the three most essential ones as *randomness*, *fuzziness* and *ambiguity*. In the world of altmetrics, all three demons, randomness, fuzziness and most notably ambiguity, may become even more crucial than in traditional informetrics. In this context we would also like to refer to the arguments by Sugimoto (2016) and Glänzel & Chi (2016) in demand for more transparency and clarity in the data covered and the need for clear definition of actors on both sides. In particular, if one talks about impact – impact upon whom is meant and, furthermore, what are the potential biases in terms of actor and user profiles? Without clarification of such issues any attempt of standardization, normalization and benchmarking of metrics would remain unsuccessful. In particular, we have, similarly to other recent studies in this paper too, analysed the correlation between altmetrics and traditional bibliometric indicators, but we did not aim at evaluating the usefulness of metrics on the basis of that correlation, nor at searching for causal relationship between the metrics under study. We expect the real (added) value of the new metrics in providing additional information that cannot directly conclude indicators of scholarly communication.

Summarising our results and observations, we can say that the example of public health has confirmed that the extension of scientometrics beyond the scope of scholarly communication remains a challenge. Significance and robustness of measures did not yet meet the standards of traditional bibliometric tools and the interpretability of altmetrics indicators requires even more context analysis than those of scholarly communication. At this moment, we find that the currently used altmetric metrics to measure the broader impact of research still fall short of the enormous expectations and the sometimes nonreflective enthusiasm in their use. Nonetheless, some of these new metrics may already provide useful information based on the feedback of broader, often heterogeneous groups of users that could be useful as supplement to traditional bibliometric indicators indeed.

Acknowledgement

This paper is an extended version and follow-up of an unpublished study presented at the COLLNET meeting in Macau on 5–8 December 2018.

References

Bookstein, A. (1997). Informetric Distributions. III. Ambiguity and Randomness. *JASIS*, 48(1), 2–10.

Brody, T., Harnad, S. & Carr, L. (2006). Earlier web usage statistics as predictors of later citation impact, *JASIST*, 57(8), 1060-1072.

Chi, P.-S. & Glänzel, W. (2018). Comparison of citation and usage indicators in research assessment in scientific disciplines and journals. *Scientometrics* 116(1), 537-554.

Chi, P.-S. & Glänzel, W. (2019), Citation and usage indicators for monographic literature in the Book

Citation Index in the social sciences. *ISSI Newsletter*, 14(4), 80–86.
Chi, P.-S., Gorraiz, J. & Glänzel, W. (2018). *Comparisons of altmetric, usage and citation indicators in chemistry disciplines.* In: R. Costas, Th. Franssen, A. Yegros-Yegros (eds.), Science, technology and Innovation indicators in transition. Proceedings of the STI Conference 2018, Leiden University, 2018, 174-183.
Costas, R., Zahedi, Z. & Wouters, P. (2015). Do "Altmetrics" Correlate With Citations? Extensive Comparison of Altmetric Indicators With Citations From a Multidisciplinary Perspective. *JASIST*, 66(10), 2003–2019.
Glänzel, W. & Chi, P.-S. (2016). Scientometrics 2.0 – and beyond? Background, promises, challenges and limitations. *ISSI Newsletter*, 12(3), 33–36.
Glänzel, W. & Thijs, B. (2011). Using 'core documents' for the representation of clusters and topics. *Scientometrics*, 88(1), 297-309.
Glänzel, W. & Thijs, B. (2012). Using 'core documents' for detecting and labelling new emerging topics. *Scientometrics*, 91(2), 399-416.
Glänzel, W. & Heeffer, S. (2014). *Cross-national preferences and similarities in downloads and citations of scientific articles: A pilot study*. In: E. Noyons (Ed.), Context Counts: Pathways to Master Big and Little Data. Proceedings of the STI Conference 2014, Leiden University, 2014, 207-215.
Glänzel, W. & Schubert, A. (1988). Characteristic Scores and Scales in Assessing Citation Impact. *Journal of Information Science*, 14(2), 123-127.
Gorraiz, J., Gumpenberger, C. & Schloegl, C. (2014). Usage versus citation behaviours in four subject areas. *Scientometrics*, 101(2), 1077-1095.Gumpenberger, Ch., Glänzel, W. & Gorraiz, J. (2016). The ecstasy and the agony of the altmetric score. *Scientometrics*, 108(2), 977-982.
Lewison, G. (2004). *Citations to papers from other documents. Evaluation of the Practical Effects of Biomedical Research*. In: H.F. Moed et al. (eds.), Handbook of Quantitative Science and Technology Research, 457-472.
Lewison, G. (2008). The returns to society from medical research (in Spanish). *Medicina Clínica*, 131 Suppl. 5, 42-47.
Moed, H. F. (2005). Statistical relationships between downloads and citations at the level of individual documents within a single journal, *JASIST*, 56(10), 1088-1097.
PlumX (2018), *PlumX Metrics*. Plum Analytics. Accessible at: https://plumanalytics.com/learn/about-metrics/.
Priem, J. & Hemminger, B. H. (2010). Scientometrics 2.0: New metrics of scholarly impact on the social Web. *First Monday,* DOI:10.5210/fm.v15i7.2874.
Sugimoto, C. (2016), *Unlocking social data for science indicator*s. (White paper), NSF Workshop on Bibliometric Indicators, Arlington.
Thelwall, M. (2018). Early Mendeley readers correlate with later citation counts. *Scientometrics*, 115(3), 1231-1240.
Thelwall, M. (2017a). Are Mendeley reader counts high enough for research evaluations when articles are published? *Aslib Journal of Information Management*, 69(2), 174-183.
Thelwall, M. (2017b). Are Mendeley reader counts useful impact indicators in all fields? *Scientometrics*, 113(3), 1721-1731.
Thelwall, M. (2012). Journal impact evaluation: a webometric perspective, *Scientometrics*, 92(2), 429-441.
Zahedi, Z., Costas, R., & Wouters, P. (2014). How well developed are altmetrics? A cross-disciplinary analysis of the presence of 'alternative metrics' in scientific publications. *Scientometrics*, 101(2), 1491-1513.

Text-Mining Historical Sources to Trace Technological Change: The Case of Mass Production

Frédérique Bone[1] and Daniele Rotolo[2]

[1] f.bone@sussex.ac.uk
SPRU, Business School, University of Sussex, Brighton BN1 9SL (United Kingdom)

[2] d.rotolo@sussex.ac.uk
SPRU, Business School, University of Sussex, Brighton BN1 9SL (United Kingdom)

Abstract

This paper explores the use of large-scale and longitudinal textual analysis of historical sources to trace technological change over periods longer than 100 years. The notion of technical change has been central to research efforts in Economics of Innovation, Science Policy and Innovation Studies, and Science and Technology Studies. Research efforts in these areas have focused on different aspects of technological change: these ranging from examining the determinants of technological trajectories on the basis of quantitative analysis of technological artifacts, to investigating the socio-technical factors shaping technological evolution on the basis of qualitative historical case studies. Building upon recent advances in text-mining techniques, this paper examines to what extent technological and social aspects can be jointly explored with the analysis of historical textual sources. To do so, the paper explores how these techniques can be used to trace technical change, to explore controversies relating to the acceptance of technologies, and to map the diffusion of technologies across socio-technical systems using the case of *mass production*.

Keywords: technical change; socio-technical systems; text-mining; controversies; sentiment analysis

Introduction and related literature

The study of technological change has been central to research efforts in innovation research. On the one hand, scholars in Economics of Innovation and Science Policy and Innovation Studies have focused on examining the role of technical change in explaining economic growth. These efforts have particularly looked at the supply and demand determinants of technical progress (Dosi, 1982; Freeman & Soete, 1997; Pavitt & Soete, 1980) and identified Research and Development (R&D) as a major contributor to technological change. Empirical analyses in this area have often relied on patent data as a proxy to examine technologies in the form of artifacts (Jaffe & Trajtenberg, 2002; Patel & Pavitt, 1997).

On the other hand, Science and Technology Studies research has paid more attention to the social aspects that underlie technological development (e.g. societal acceptance). Empirical contributions in this area also focused on the understanding of the development of socio-technical regimes, i.e. set of rules "*embedded in a complex of engineering practices, production process technologies, product characteristics, skills and procedures, ways of handling relevant artifacts and persons and ways of defining problems - all of them embedded in institutions and infrastructures"* (Rip & Kemp, 1998), and on transitions from one regime to another. Empirical evidence in support of these theories has been mainly provided through historical narratives and case studies (Geels, 2002; Geels & Schot, 2007).

Although these studies have provided important contributions to the understanding of technological change, increasing access to large-scale textual sources in a digital form as well as advancements in text-mining techniques have provided the opportunity to expand on previous studies along two main directions: (i) the analysis of the full-text of documents, thus

extending the object of analysis from technological artifacts to other types of socio-technical processes (e.g. process innovation, organisational innovation, social acceptability of technologies); (ii) the examination of much longer periods of observation.

This paper explores the use of large-scale and longitudinal textual analysis of historical sources to trace technical change over long periods (over 100 years). Moving from using historical textual sources for providing a qualitative historical analysis to a quantitative account of socio-technical change poses a number of challenges that will also be discussed. In order to exemplify these challenges, the paper focus on the study of a specific case, that is *mass production*, using the full-text data of "Scientific American" journal from 1845 to 2005. *Mass production* is here defined as methods of production which enables high volumes of production to produce standardised products (Womack, Jones, & Roos, 1990). This paper specifically explores the technical aspects of using general science and technology magazines to trace technological change.

Method and data

This section focuses first on the rationale for selecting Scientific American as a source of textual data for our analysis of the *mass production* case; it then describes the process of data collection and text extraction together with a brief overview of how textual data will be analysed.

Textual data: Selection and content evolution of the source

The selection of an appropriate textual source has to be grounded on both theoretical and practical aspects of data availability and collection. In terms of theoretical aspects, the data source needs to match-up with the technology one aims to trace; in the case of *mass production*, the source needs to have temporal continuity over a period of a century. The source should also enable to observe diffusion of technology beyond a single sector or socio-technical system. A third aspect concerns the acceptance of a given technology: to what extent the source enables us to map debates around a given technology? Finally, in practical terms, the source should be accessible online in a digital format over a long period of time, thus facilitating the processing of the large amount of text data.

In this paper, several sources used for historical analysis has been considered. These include trade journals, technical magazines, general science and technology magazines, conference proceedings of engineering conferences, and generalist newspapers. We could not identify any single source that can fully meet all the aspects elaborated above, but some sources offer more potential than others. For instance, trade journals do not offer the possibility to study diffusion of technologies across sectors and/or socio-technical systems. Engineering magazines are limited by the access and digitalisation of sources even when their coverage of the technological paradigm is relatively broad. In terms of coverage and potential to study diffusion, more general sources such as technology and general science magazines and newspapers provide more ground to such analyses.

The two differentiating aspects to consider are (i) the trade-off between the study of controversy and (ii) the ability of the resource to consistently discuss the selected technology over time. While newspapers data offer a better standpoint at studying controversies, and especially the aspect of the acceptance of technologies by the general public, they offer a limited opportunity to study the evolution of technologies. Finally, general science and technology magazines are more likely to use a consistent terminology to describe technologies, which makes it a better candidate to identify relevant articles through a set of expert-defined keywords.

The Scientific American journal has been selected because it is a source which discusses over time and at length new inventions and technologies. Its use may have limitations in terms of understanding controversies, mostly in terms of reservations that the general public may have over a technology.

The content of Scientific American has been subject to a number of changes over the years through both variations in editorship or significant changes in the journal format. The analysis and for interpretation purposes should reflect on these changes. Scientific American is a journal that has focused on the developments of science and technology since 1845. The journal has known two main phases: the first one from 1845 to April 1948 when the journal focused mainly on technology and a second one from May 1948 to present, when the journal shifted focus towards the popularisation of scientific breakthroughs (Figure 1 depicts how the corpus of the source evolved in the light of major editorial changes).

Data collection and extraction

The textual data of Scientific American were downloaded from Nature's website, representing 16,868 PDF files. There are different file types (i.e. articles and issues) according to the observation period due to changes with the digital format of the journal. Each PDF document was processed with an Optical Character Recognition (OCR) software (i.e. Omnipage Ultimate) which retrieved the textual parts of the document, and preserved the structure of sentences as structured in columns. Textual data were then prepared for text-mining removing hyphenation, line breaks, numbers, punctuation, non-graphical characters, multiple spaces, and also converting the text to lower case.

Data analysis

The study focuses on identifying threads of discussion surrounding the case of *mass production* as reported in the corpus of Scientific American. To do so, a list of keywords associated with *mass production* was compiled on the basis of a review of the literature and with the support of experts. Keywords and regular expressions enabled us to identify sections of the corpus related to *mass production*[1]. We examined these sections using text-mining techniques with the support of the *tidytext* package in R (Silge & Robinson, 2017). We first analysed the evolution of the whole vocabulary of Scientific American in terms of average number of words per issue, yearly total number of words, and yearly total number of distinct words. This enabled us to increase our understanding of how the corpus of Scientific American has evolved over 150 years of observation, thus providing a base for comparison and normalisation. We then examined longitudinally the occurrence and co-occurrence of keywords to generate evidence of how the debate around *mass production* has intensified or declined on specific areas/themes over the observation period. We concluded the analysis performing a sentiment analysis of the textual data around each keyword (we extracted 50 words before and after each keyword occurrence), as a proxy of acceptance of the technology. This was assessed using the Bing lexicon available in the *tidytext* package in R (Silge & Robinson, 2017). The results of these analyses are presented in the next Section.

[1] Keywords and regular expressions used can be found on https://github.com/Frederique85/Deep-Transitions/blob/master/Mass_Production-Kwslist_regex.csv .

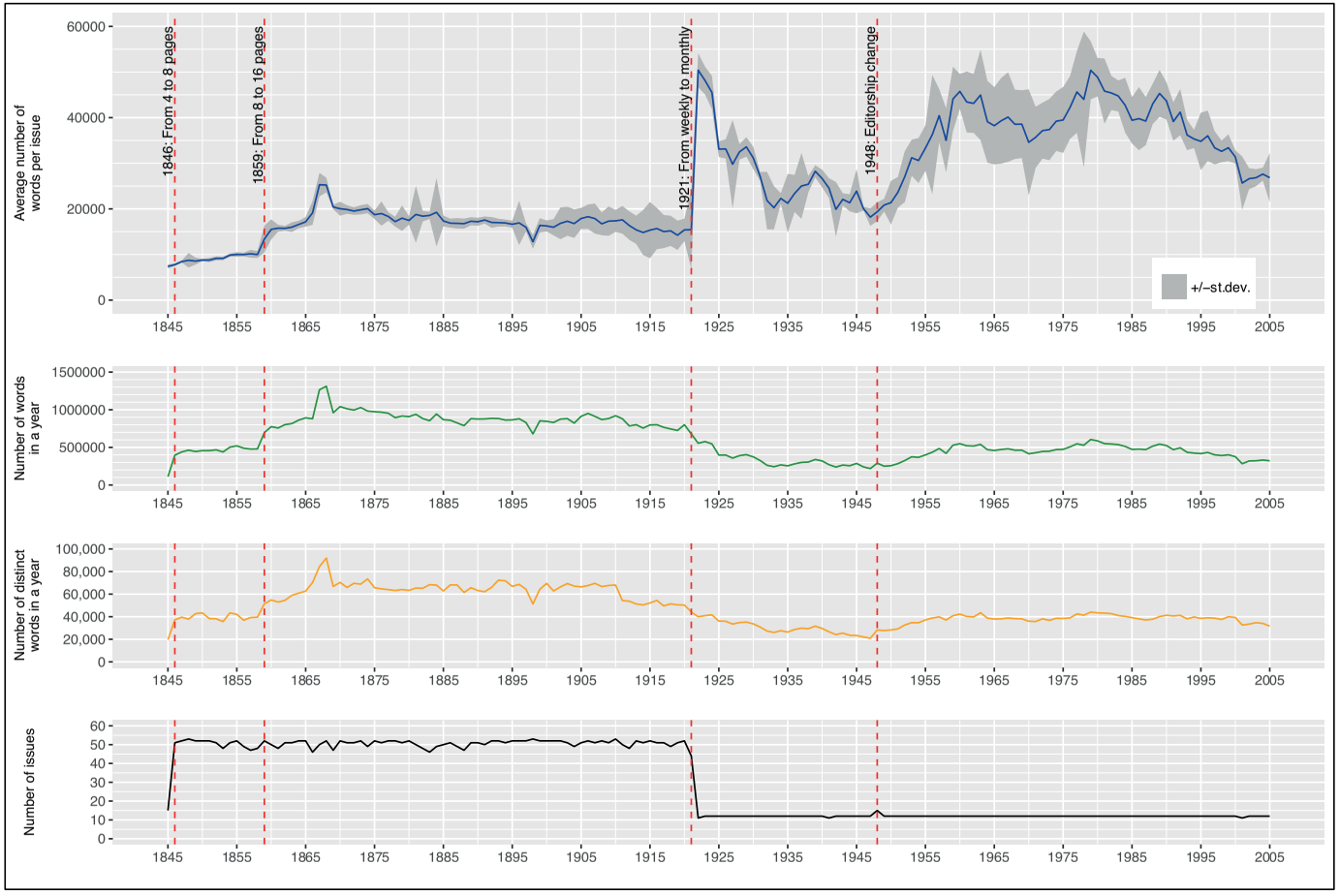


Figure 1: The evolution of the corpus of the Scientific American (1845-2005)

Results

Data coverage: How does the evolution of the source affect textual data?

We examined the corpus of Scientific American in terms of average number of words per issue, yearly total number of words, and yearly total number of distinct words. These indicators provide evidence of how the corpus has changed in terms of length (as proxied by the average number of words per issue and the yearly total number words) and vocabulary (as proxied by the yearly total number of distinct words).

Figure 1 illustrates the results of this analysis with the four major editorial changes of the source we discussed in the previous section. These changes considerably affected the corpus of Scientific American. In 1846, the change of the length of the issues from four to eight pages generated a rapid increase of the length of the corpus – from a yearly total number of words of about 110,000 to 390,000-520,000. In 1859, issues were further expanded to 16 pages, thus increasing the yearly total number of words to about 880,000 with peaks of 1,000,000-1,300,000 words from 1867 to 1873. With the change of Scientific American from a weekly to a monthly journal in 1921, we observed a decline of the yearly total number of words to a minimum of about 218,000 words in 1947, followed by a rapid growth to 605,000 words after the editorship change in 1948. After 1921, we also observed a rapid increase of the average number of words per issues as a result of the lower number of issues per year (from 52 to 12 issues per year). Similar trends are observed when considering the size of the vocabulary of words used in Scientific American articles: from about 20,000 words to 36,000-43,000 words in 1846, to a maximum of about 92,000 until 1921, to a vocabulary ranging from 21,000 to 44,000 words per year in the subsequent period. We will account for these changes of the size of the corpus for normalisation and interpretation purpose.

Mapping mass production: Keywords occurrence and co-occurrence

We used regular expressions to identify the total number of occurrences for each keyword in the corpus of Scientific American from 1845 to 2005. The results of this analysis are reported in Figure 2 for the whole period of observation, while Figure 3 shows the occurrence of the top 20 most frequent keywords by year. Keywords were identified by experts.The preliminary results show a number of trends. The term *interchangeability* (and its variations) is the most frequently occurring word and is consistently mentioned over the whole observation period, with two peaks: one around 1900 and World War II (1945). The keywords related to *robots* are also very frequent, but its appearance is concentrated in a much lower number of issues, i.e. about 10% of the issues. Figure 3 shows that these keywords are frequently mentioned after 1945 with a sharp increase after 1970. The keywords *mass production* and *mass manufacturing* are peaking in different periods, 1845 and 1920, respectively. This seems to suggest a change in the vocabulary.

Mapping mass production: Sentiment analysis

Figure 4 explores the amount of positive and negative words co-occurring with the top 20 most frequent keywords. For each keyword, bars depict the number of positive words (above the red line at 0) and negative words (below the red line at 0) in a given year. Figure 4 provides insights on whether keywords are talked about in a more positive or negative light. Below are some observations about trends observed. These preliminary results provide evidence of a number of trends. The term *flexible manufacturing* seems to be seen mostly in a positive light. *Robots* and *robotics* seem to be mostly controversial especially around 1998-2005.

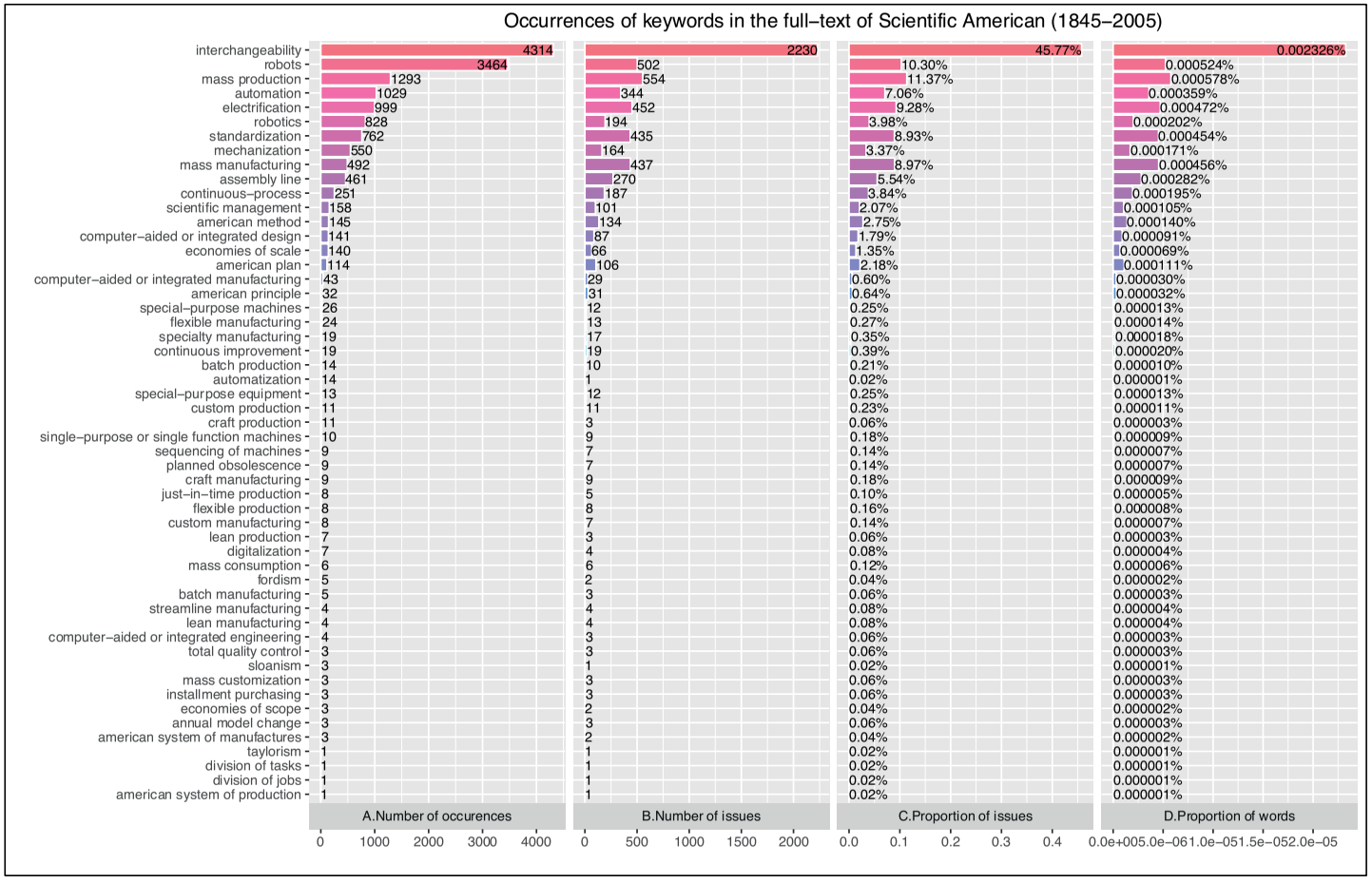


Figure 2: Keyword occurrence in Scientific American (1854-2005)

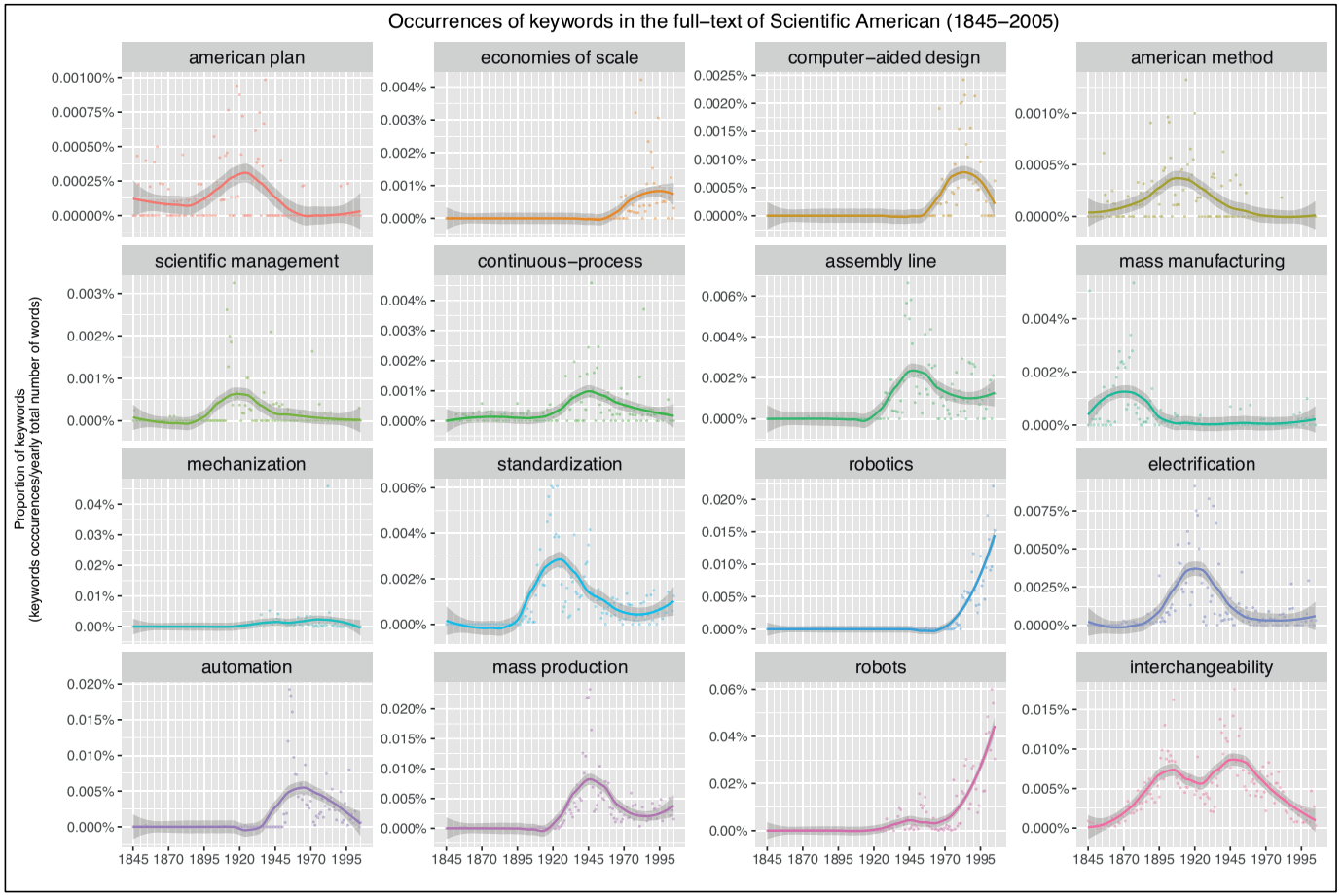


Figure 3: Occurrence of the top 20 most frequent keywords in Scientific American (1945-2005)

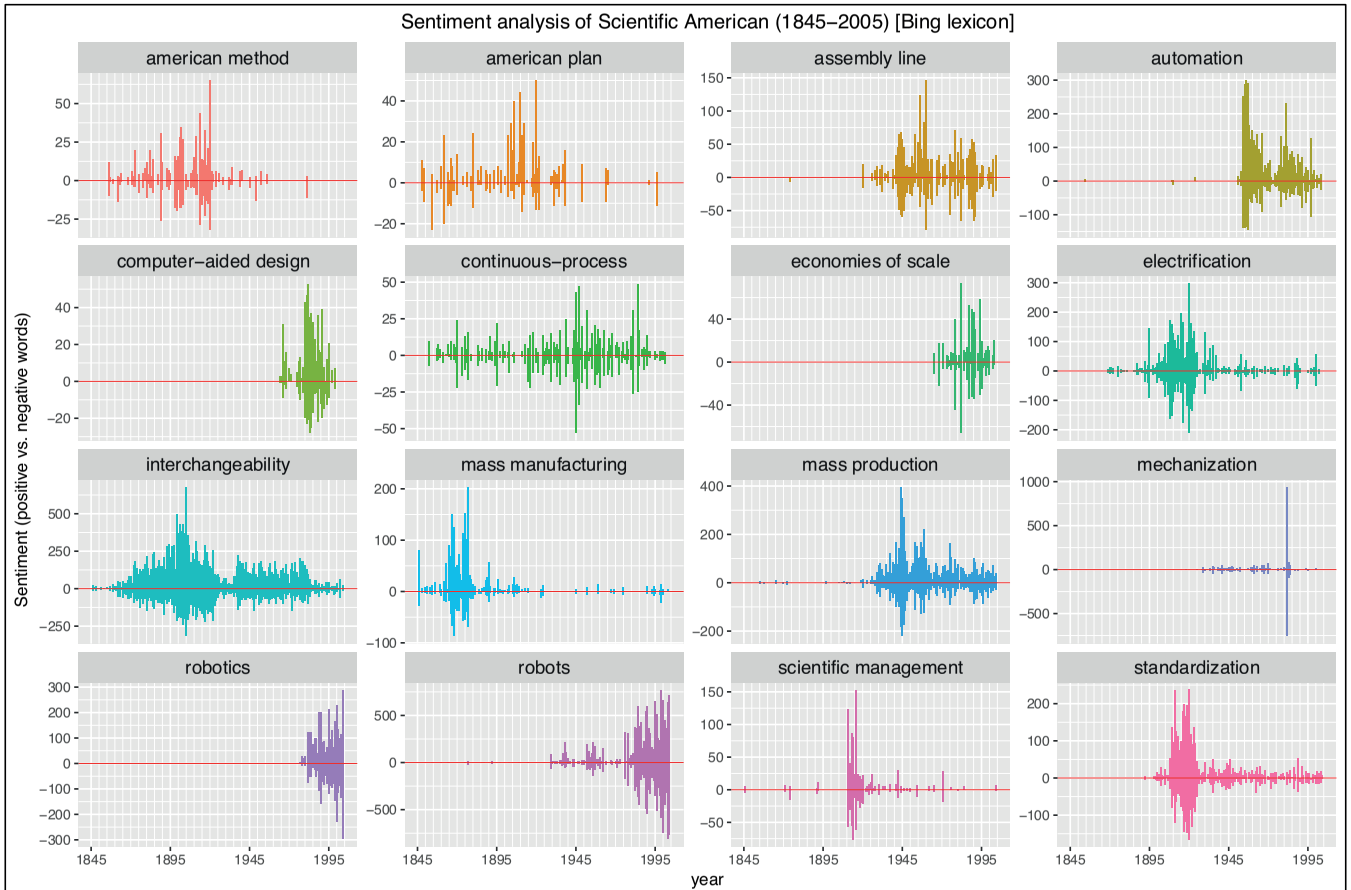


Figure 4: Sentiment analysis of Scientific American (1845-2005)

American principles, *continuous process* and *economies of scale* seem also attracting a higher share of negative comments. *Interchangeability* seems most controversial in the early 1900s, but in the second wave of discussion it is seen in a positive light. While the sentiment analysis provides a first insight of whether some keywords are used in a positive or negative discussion, this does not consider whether the general tone of the journal is using a more or less positive tone overall. In general, written corpora are using more positive than negative words (Augustine, Mehl, & Larsen, 2011; Dodds et al., 2015; Iliev, Hoover, Dehghani, & Axelrod, 2016). In order to make sure our analysis is not affected by this bias, the paper explores the Language Positive Bias (LPB) of the journal over time. This is depicted in Figure 5. The upper chart, Figure 5a, shows the overall language bias of the source over the overall period. In the early days of the journal, the level of the positive bias is relatively high, around 2. From 1846 to 1867, it decreases from 2 to about 1.5. Over subsequent years until 1911 the bias is more or less stable around 1.7. Between 1911 and 1936, the bias is again stable around 1.5. The changes of editorship seem to coincide with a big drop in positive bias. After 1956 the positive bias has generally been decreasing to reach a very moderate positive bias.

When comparing the positive bias to the ways that the two mostly occurring keywords have been discussed, one may interpret the results slightly differently (Figure 5b and 5c). It seems that *interchangeability* has been discussed following the overall line of the journal. Though looking at individual points, from 1845 to 1967 there is a lot of variation about how it is talked about across the years, and some years it was talked about quite negatively. In the 1890s, the outliers show that the tone was more positive than the tone of the journal. Between 1945 and 1995 *interchangeability* is seen in quite a positive manner.

When looking at *robots*, from 1855 to 1910, the keyword was generally seen in a negative manner. However, at this stage as the previous analysis showed, the term *robots* were not used very frequently. Between 1915 and 1940, it was seen in a more positive light. While following the general tone of the journal thereafter, in the 1970s it was also seen more positively compared to the journal's tone. These results have to be contrasted with our previous results. While *robots* can be seen to be discussed more negatively than other concepts, this effect may be partly due to the fact that the journal over more recent years had a more negative tone. However, Figure 5c still shows that robots are seen as controversial since the 1990s.

Future steps

The analysis presented in this paper gives early insights on how the discourse of *mass production* has evolved over time. It has first shown that not all keywords are used equally over time, topics related to interchangeability and robots are much more prominent than others. The results also show that for many keywords, there is a `hype' period (see Figure 3) at which they are more frequently discussed, which is in general isolated in a specific time period. Further analysis still needs to be performed to check the results. There needs to be further work on the development of keywords, and the analysis on keywords could be grouped in broader issues to enable a clearer analysis of trends.

There are also limitations to this work. The paper focus on exploring trend of *mass production* looking at a single source. The results presented may be subject to editorial bias, and there is a question whether these could be replicated using other sources. There are also some limitations in the sentiment analysis. The lexicons of positive/negative words are validated on styles of writing that are considerably different from our textual sources (Silge & Robinson, 2017). Further refinement of these dictionary will be needed to improve the accuracy of the analysis.

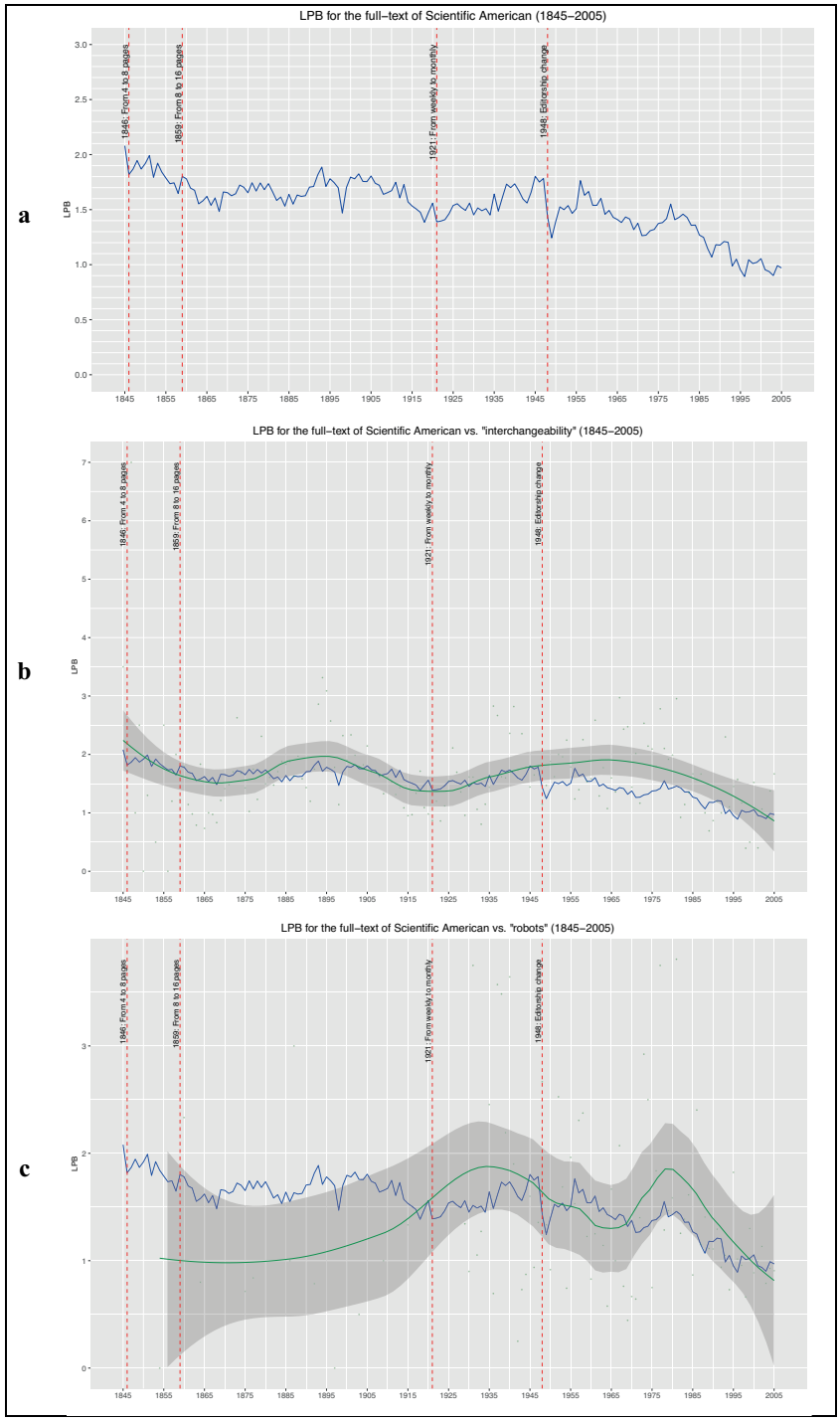


Figure 5: Language Positive bias of Scientific American

Acknowledgments

The authors wish to thank James Anderson and Baillie Gifford & Co for their support.

References

Augustine, A. A., Mehl, M. R., & Larsen, R. J. (2011). A positivity bias in written and spoken English and its moderation by personality and gender. *Social Psychological and Personality Science*, *2*, 508–515. https://doi.org/10.1177/1948550611399154

Dodds, P. S., Clark, E. M., Desu, S., Frank, M. R., Reagan, A. J., Williams, J. R., … Danforth, C. M. (2015). Human language reveals a universal positivity bias. *Proceedings of the National Academy of Sciences*, *112*(8), 2389 LP-2394. Retrieved from http://www.pnas.org/content/112/8/2389.abstract

Dosi, G. (1982). Technological paradigms and technological trajectories. A suggested interpretation of the determinants and directions of technical change. *Research Policy*, *11*(3), 147–162. https://doi.org/10.1016/0048-7333(82)90016-6

Freeman, C., & Soete, L. (1997). *Economics of Industrial Innovation*. *Simpósio GIT*. Cambridge, MA: MIT Press.

Geels, F. W. (2002). Technological transitions as evolutionary reconfiguration processes: a multi-level perspective and a case-study. *Research Policy*, *31*(8–9), 1257–1274. https://doi.org/10.1016/S0048-7333(02)00062-8

Geels, F. W., & Schot, J. (2007). Typology of sociotechnical transition pathways. *Research Policy*, *36*(3), 399–417. https://doi.org/10.1016/j.respol.2007.01.003

Iliev, R., Hoover, J., Dehghani, M., & Axelrod, R. (2016). Linguistic positivity in historical texts reflects dynamic environmental and psychological factors. *Proceedings of the National Academy of Sciences*, *113*(49), E7871–E7879. https://doi.org/10.1073/PNAS.1612058113

Jaffe, A. B., & Trajtenberg, M. (2002). *Patents, Citations, and Innovations: A Window on the Knowledge Economytle*. Cambridge, MA and London, UK: MIT Press.

Patel, P., & Pavitt, K. (1997). The technological competencies of the world's largest firms: Complex and path-dependent, but not much variety. *Research Policy*, *26*(2), 141–156. https://doi.org/10.1016/S0048-7333(97)00005-X

Pavitt, K., & Soete, L. (1980). Innovative activities and export shares: Some comparisons between industries and countries. In K. Pavitt (Ed.), *Technical Innovation and British Economic Performance* (pp. 38–66). London, UK: Palgrave Macmillan. https://doi.org/10.1007/978-1-349-04561-7

Rip, A., & Kemp, R. (1998). Technological change. In S. Rayner & E. L. Malone (Eds.), *Human Choice and Climate Change* (pp. 327–399). Columbus, Ohio: Battelle Press. https://doi.org/10.1007/BF02887432

Silge, J., & Robinson, D. (2017). *Text mining with R: A tidy approach*. Sebastopol, CA: O'Reilly Media.

Womack, J. P., Jones, D. T., & Roos, D. (1990). The Machine that Changed the World: The Story of Lean Production. *World*. https://doi.org/10.1016/0024-6301(92)90400-V

Measuring the Impact of an Author of Multi-Authored Articles - Aggregating Metrics for Multiple Authors' Analysis

D. Gnana Bharathi

dgbharathi@gmail.com
Environmental Science & Engineering Division
Central Leather Research Institute, Chennai, India - 600020

Abstract

Methodologies for evaluation of scientific articles written by multiple authors can be divided into two categories: equal sharing and differential sharing of citations among the coauthors. The proposed aggregating metrics for multiple authors' analysis (*ammaa*) shares the fractions of total citations equally among the coauthors. A threshold limit which is a multiplier of co-authorship is introduced. For any article, the total citations are squared and are divided by the threshold limit and the number of authors. This results in an increase in impact, for every author, which is more than the proportional division of total citations by a number of authors. This increased impact continues until it reaches the citation limit set by the threshold limit for the coauthors. It ensures that every author equally earns improved credit with an increase in citations until the citations of the article reach the limit. At this threshold limit, *ammaa* merges with total citations, and from that point, every author gets full credit of all citations. This method can be extended to other evaluation processes, such as altmetrics, where the number of views, number of sharing, number of downloads, etc. are measured. .

Introduction

Evaluation of research publication plays an important role for the authors for their research findings (Waltman, 2016). Indices such as h-Index, g-Index, I10- Index, etc. are used for the evaluation of individual researchers (Vieira & Gomes, 2011). However, these methods do not differentiate whether the articles are written by a single author or multiple authors. As a result, every author claims full credit for the article, irrespective of the number of authors. Advancements in science and technology necessitate collaboration among researchers from many fields (Cummings & Kiesler, 2005; Gazni, Sugimoto, & Didegah, 2012; Roux, Rogers, Biggs, Ashton, & Sergeant, 2006; Seddon, Armstrong, & Maloney, 2007). However, there exists a trend of increase in coauthorship by including those whose contributions are categorized as "guest author" (Wager, Singhvi, & Kleinert, 2015) , "gift author" (Jack, 2015), "nonacademic collaborator" (Sarna-Wojcicki, Perret, Eitzel, & Fortmann, 2017), etc. In scientific journals, it is increasingly becoming normal that almost every article is written by more than one author (Cronin, Shaw, & La Barre, 2003; Leimu & Koricheva, 2005; Shaban, 2007). However, all the authors getting 100 percent credit for every citation for the coauthored articles has always raised objections (Cooper, 2003; Figg et al., 2006).

Coauthorship and impact

Research findings are as a result of an effort by an individual or group of researchers. In many circumstances, authors have to collaborate with other peers as the works involve multidisciplinary subjects(Porter & Rafols, 2009). Increased cooperation may or may not result in major research finding. A number of articles published by multiple authors fail to get impact in terms of citations, downloads, etc. (Iribarren-Maestro, Lascurain-Sanchez, & Sanz-Casado, 2007). As the coauthors are increasing, in recent times, the journals are asking for specific contribution by each author (Ilakovac, Fister, Marusic, & Marusic, 2007; Tarnow, 2002; Yager, 2007). Despite these issues, practically there is no restriction on a number of authors.

Eight different authors and their cited articles, many of which co-authored, are shown in Fig. 1 & Fig. 2. The pattern of citations, as well as coauthorship of the articles, show different trend. Irrespective of whether the article is single authored or co-authored, some articles get poor citations while others get more citations. Year of publications too plays an important role in the increase in citations. However, not all the articles will keep cited for a long duration.

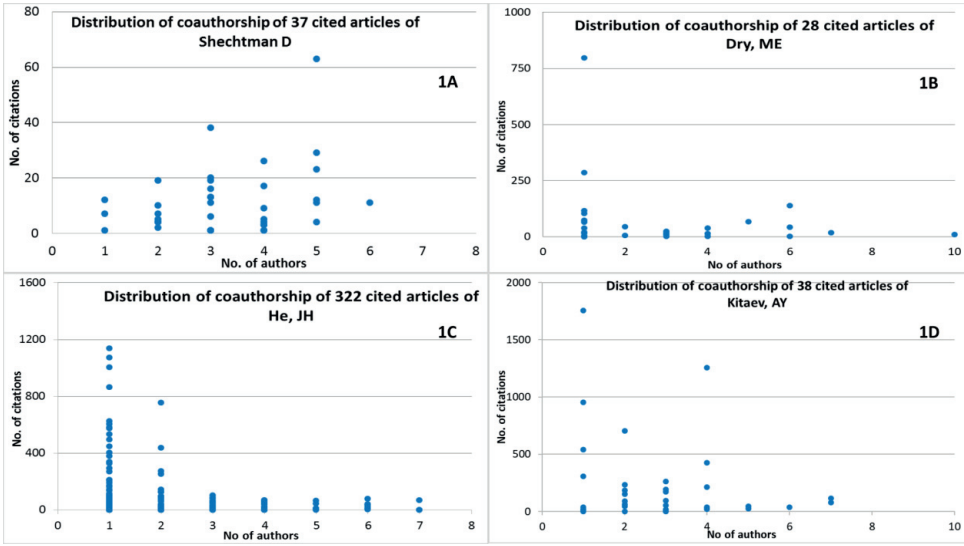


Figure 1. Distribution of citations for the articles published by the selective researchers who published articles with limited number of co-authors.

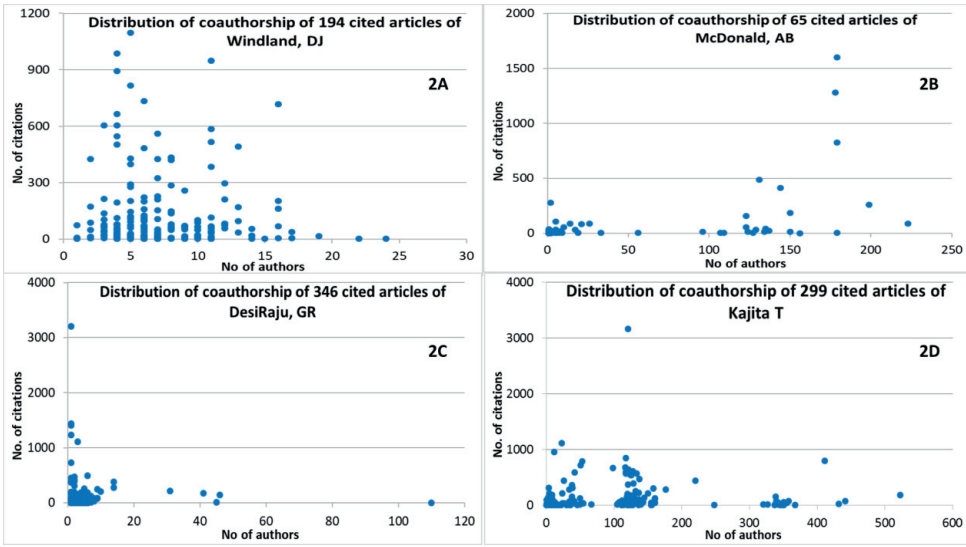


Figure 2. Distribution of citations for the articles published by the selective researchers who published articles with more number of co-authors.

Among the researchers, *Shechtman* gets more citations with an increased number of co-authors (Fig 1A). On the other hand, single-authored articles by authors like *He* and *DesiRaju* have more citations than the co-authored articles (Fig 1C & Fig 2C). Publications of the articles by *He* are mostly single-authored ones with the maximum number of coauthors being seven (Fig 1C). *DesiRaju* published many of his articles with a very limited number of coauthors. Only a few of his articles are published with more coauthors (Fig 2C). Articles such as those co-authored by *Wineland* and *Kajita* show no specific trends (Fig 2A & 2D). Authors like *Mc Donald* and *Kajita* have published articles with more than hundreds of coauthors (Fig. 2B & 2D).

These patterns clearly indicate that giving full credit to every coauthor will undermine the researchers who published with a fewer number of coauthors or as a lone author. As a result, when the citations or downloads are counted for evaluation, it is essential to study the impact of coauthorship on these parameters carefully. For the articles of a given number of citations, the due impact should be given for the single-authored ones in comparison with coauthored ones. Further, among the coauthored articles, those with less number of authors should get more impact for a given number of citations.

There are a number of methods to assess the impact of coauthored articles. They can broadly be classified as equal sharing of citations and selective sharing of citations. There are several methods for the evaluation of the selective sharing of citations, such as higher impact for the first author, first two authors, senior author, etc. (Ausloos, 2013; Egghe, Rousseau, & Van Hooydonk, 2000; Sekercioglu, 2008; Shen & Barabasi, 2014; Tol, 2011; Xu, Ding, & Malic, 2015). Tscharntke, Hochberg, Rand, Resh, & Krauss, (2007) proposed various methods to evaluate the impact of the coauthors. The equal sharing of citations can be further classified as fractional sharing of total citations and full sharing of every citation (Galam, 2011; Hagen, 2010; Liu & Fang, 2012). The last one is the most common practice. However, these methods have limitations on a wider scope. A new method that simplifies the evaluation of collaboration of researchers and also for measuring the share of the impact is proposed.

Calculation of *ammaa*

The new method to evaluate multiple authors is called as "aggregating metrics for multiple authors' analysis" (*ammaa*). The method divides the square of citations by a number of authors to provide an impact for each author. In this metric, all authors have equal sharing and on every addition of citations, the percentage of their share increases. This metrics works similar to the performance incentives provided for every increase in the productivity on an exponential scale until a benchmark is achieved. Once, the target is achieved, the all members of the team get postings with each member getting full salary.

Threshold limit (T), a multiplier of co-authorship, in terms of a number of citations, is introduced. For any coauthored article, T multiplies with the number of authors (a). As a result, the increase in coauthors necessitates increased citations for the increased ownership of citations. All coauthors can claim full credit for total citations, once the number of citations is $(aT-T)$ or more.

The value of T can be set by universal consensus or specific to a country, group, or organizations, as well as by subject wise, institution wise, or group wise. T can also be assigned to the number of downloads, number of sharing, number of views, etc. T can be 10, 15, 100, 1000 or less or more. Subjects that traditionally receive fewer citations such as geology, mathematics, etc. may have low T. Subjects that get high citations such as biochemistry, biotechnology, etc. may have high T. For universal standard, T may be fixed at

100 for any subject. Any other value can also be set based on consensus. Once *T* is set, *ammaa* can be calculated for every author of any multi-authored article. The calculation for *ammaa for an a*rticle is as follows:

For single-authored articles, *ammaa* equals total citations

For any coauthored articles

$$ammaa = \frac{\left(c + \left(\frac{c^2}{T}\right)\right)}{a}, where\ ammaa \leq (\boldsymbol{aT} - \boldsymbol{T}) \qquad - I$$

Otherwise

$$ammaa = total\ citations \qquad - II$$

c= Total citations
a= number of authors
T= threshold limit

Methodology
The equation (I) is derived as follows: the square of total citations is divided by the threshold limit, *T*. This value is added with total citations. Trend lines for threshold limits 10, 25, 50, 100, 250, 500 and 1000 are shown in Fig 3 as *citations Vs. (c+(c^2/T))*. These trend lines form the basis for the calculation of *ammaa*.

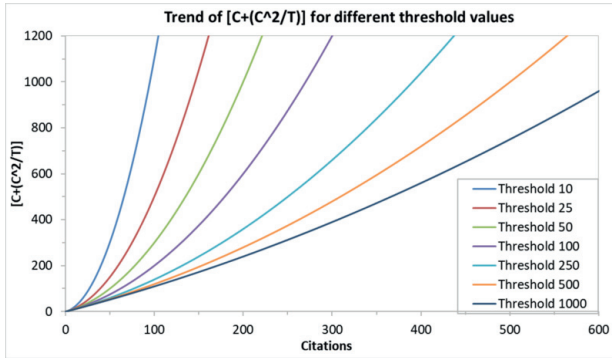


Figure 3. Trends of (c+(c^2/*T*)) for the citations for different threshold limits, *T*.

For "*a*" number of authors, which receives "*c"* number of citations, *ammaa* for each author is (c+*(c^2/T*)) divided by the "*a*" number of authors. For a threshold limit of 100, *ammaa* for 2, 3, 4, and 5 authors, along with the trend line for *T*=100, is shown in Fig 4. Similarly, *ammaa* can be applied to any number of citations of the article authored by any number of coauthors. The endpoints merge with the total citations which are clearly shown in Fig 5. In this figure *ammaa* for *T*=10 and *T*=100 are calculated. For any given number of coauthors, it follows a trend, and it merges with the total citations at a point **(*aT-T*)**. From this point onwards *ammaa* for each author is same as total citations, and every author can claim full credit for all citations received by the article.

Let us assume T is 100. For a single-authored article, every citation belongs to the author. When an article is written by two authors, citations are equally divided among the coauthors. The point of intersection between total citations and (aT-T) will be at 100th citations. As per the calculation of *ammaa*, when an article gets its first citation, each author's merit is just 0.5, precisely 0.505.

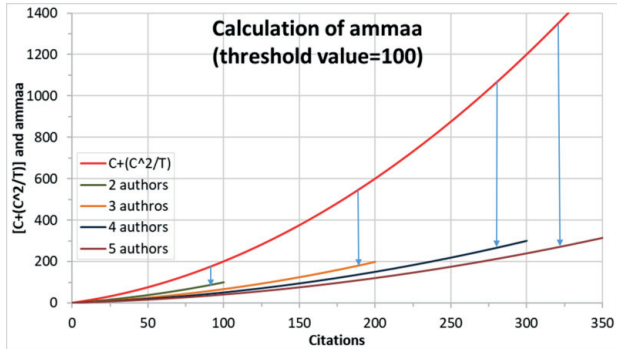


Figure 4. Calculation of *ammaa* for set of coauthors. Increase in a number of coauthors doubles the number of citations needed to achieve the merits of total citations.

When the number of citations reaches 10, the merit of each author will be 5.50, and when the number of citations is 50, the merit of each author will be 37.50, and so on. When the number of citations of the article reaches 100, merit for both authors will be 100. This demonstrates that for every addition of citation, the share of every author increases gradually, and at the same time equally. Beyond the threshold limit, *ammaa* for every author is same as if the article was authored by a single author.

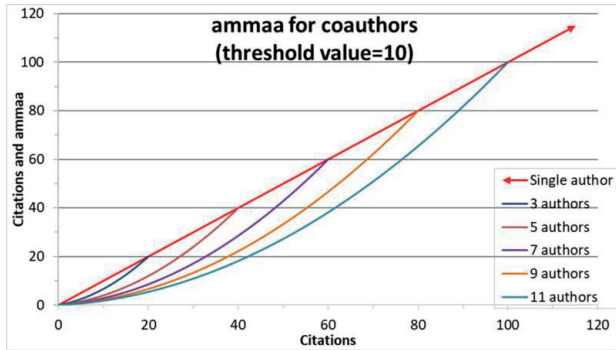


Figure 5. Calculation of *ammaa* for a different number of coauthors and the way it merges citations as by a single author. T=10 for Fig 5A and T=100 for Fig 5B.

Let us assume T is 100 and the article is written by three authors. The point of intersection between total citations and (aT-T) will be at 200th citations. For the first citation, each author's merit is 0.34. When the citations reach 10, each author's merit is 3.67, when citations are 100, each author's merit is 66.67, and so on. Once citations reach 200, all three authors merit full credibility of total citations.

Assuming T is 100, if there are four authors, the number of citations needed for full credit of all citations is 300. If there are eleven authors, the number of citations needed for full merit is 1000 and so on. Let us take some examples. If *T* is 100, and if there are six authors and the article received 53 citations, *ammaa* for each author is calculated as follows:

(*aT-T)* = (6*100-100) = 500)

$$ammaa = \frac{\left(53 + \left(\frac{53^2}{100}\right)\right)}{6}, where\ ammaa \leq 500$$

ammaa for each author is 13.52. When the article gets 100, 150, 200, and 300 citations, *ammaa* for each author is 33.33, 62.50, 100.00 and 200.00 respectively. As the citations reach 500, *ammaa* for each author will be 500.00, as *ammaa* merges with total citations. If the article continues to get more citations, *ammaa* equals total citations, and each author is credited for all the citations.

Let us say *T* is 15 for a particular subject or an institution. If an article is coauthored by six authors and the article received 53 citations, *ammaa* for each author will be 40.04 as
(*aT-T)* = (6*15-15) = 75)

$$ammaa = \frac{\left(53 + \left(\frac{53^2}{15}\right)\right)}{6}, where\ ammaa \leq 75$$

Let us say *T* is 1000 for an award or a membership of a society. If an article is coauthored by six authors and if the article received 53 citations, *ammaa* for each author will be 9.30 as

(*aT-T)* = (6*1000-1000) = 5000)

$$ammaa = \frac{\left(53 + \left(\frac{53^2}{1000}\right)\right)}{6}, where\ ammaa \leq 5000$$

Let us say *T* is 10 and an article coauthored by six authors received 53 citations.

(*aT-T)* = (6*10-10) = 50)

As 53 is greater than (*aT-T*), *ammaa* for each author will be 53.

In this case; the Eqn-I is not valid. Eqn-II should be applied. If the Eqn-I is inadvertently applied *ammaa* for each author will be calculated as 55.65 which is incorrect.

Table I shows a total number of articles written by the eleven researchers. Only those articles that received at least one citation are considered for the calculation of *ammaa*. While *Dry* has published 28 articles, *DesiRaju* published 346 articles. In general, an average number of coauthors vary widely. On an average number of coauthors for *He* is as low as 2.07 and for *Kajita* it is as high as 68.06.

Table I. Calculation of *ammaa* for different *T* values for six Nobel laureates and five scientists from BRICS countries. Data from Clarivate Analytics' Web of Science were collected in [^]May 2016, [*]September 2016 and [!]October 2016.

Name of the author	*Total cited articles*	*Total co-authors*	*Avg coauthors*	*Total citations*	*Citations / authors*	*Citations * authors*	*ammaa 10*	*ammaa 25*	*ammaa 50*	*ammaa 100*	*ammaa 1000*
Kajita, T[^]	299	20349	68.06	28551	1394	2904644	17554	12524	8415	5380	1795
Tsallis, C[*]	259	1484	5.73	15381	8647	1669837	13637	12825	12004	11054	9098
McDonald, AB[^]	65	3651	56.17	6553	363	941294	3938	1937	1222	862	438
Wineland, DJ[!]	192	1375	7.16	25066	4662	173817	24238	22711	20646	17209	6794
Haroche, S[^]	124	705	5.69	16430	3421	104171	15937	15077	14092	12517	5518
DesiRaju, GR[*]	346	1504	4.35	26976	15862	99387	26126	24315	22337	20496	16935
Edwards, RG[!]	253	1036	4.09	8664	2668	47856	8017	6707	5578	4550	2907
He, JH[*]	322	668	2.07	19389	16281	28587	19172	18720	18234	17859	16808
Kitaev, AY[*]	38	114	3.00	8195	5184	18830	8161	8032	7925	7749	6000
Dry, ME[*]	28	86	3.07	1975	1630	3700	1934	1852	1756	1693	1636
Shechtman D[^]	37	122	3.30	432	139	1591	348	247	193	166	142

Total citations for these researchers also vary with *Shechtman* having 432 citations and *Kajita* with 28551 citations. Microsoft Excel software was used to calculate *ammaa* for every article authored by the eleven researchers. For every article, the values of *ammaa* for these authors were summed up. As there are differences in number articles, a number of coauthors and number of citations, *ammaa* for every researcher was calculated for different T values viz. 10, 25, 50, 100 and 1000.

Clarivate Analytics' Web of Science Core Collection database was used for the calculations. These authors have single-authored articles as well as co-authored articles. Publications by six Nobel laureates, four in the field of physics, one each in the field of chemistry and medicine were tested. Similarly, five scientists from BRICS nations who have highest citations as a single author, in their respective country, were analyzed. The Table also includes an average number of coauthors for these eleven researchers, a division of the number of citations by the number of coauthors and the total number of citations as claimed by all the coauthors.

Discussion

An author who writes articles individually gets credit for every citation from every article – be it one, hundreds or thousands. However, many researchers coauthor their publications. For all the articles published by an author, who may have coauthored some or all articles, *ammaa* can be calculated. Articles published by some researchers were tested for the applicability of the *ammaa*. (Table I).

It is well known that the number of coauthors varies based on subjects, even sub-subjects, etc. The field of neutrino, for which Kajita T and McDonald AB received Nobel Prize in 2015, needs a large number of collaborative studies for the successful publication. As a result, for Kajita, if the citations of each article divided by the number of coauthors, the citations for the author will be just 1394. This is far below the total citations, 28551. If we calculate the citations claimed by all the coauthors, the total citations will be 2904644. This is more than 100 times of total citations. As all the authors getting full credit for all the citations will be inappropriate, *ammaa* can be used for sharing the citations.

The first thing is fixing the Threshold limit, *T*. As stated earlier, any value can be fixed as *T*. The impact *T* can be visualized in Fig 6. where actual citations for each article authored by four researchers are displayed. These graphs also show citations per author, a total of citations claimed by all the authors, *ammaa 10* and *ammaa 100*. For Kitaev, as number citations are more and a number of coauthors are less, *ammaa 100* merges with total citations for some articles and most of *ammaa 10* merges with total citations. On the other hand, for McDonald, there is a distinct difference between the total citations and other values.

Table I provides five threshold limits viz. 10, 25, 50, 100 and 1000. It is clear as the field neutrino demands more coauthors. If *T* is set at 25, the credit of citations for Kajita reduces by more than 56%, McDonald loses by 70%. Therefore, subjects like neutrino *T* value may be as low as 25 or even less.

In the field of quantum computing, for which Haroche S and Wineland DI received Nobel Prize in 2012, *T* value can be 100. Scientists from BRICS countries, as shown in the Table I, have a number of single-authored publications. As a result, the difference between total citations and *ammaa* is less for them, even when *T* is 1000.

Therefore, a fixed value of 100 citations may be set as a threshold limit *T* for universal calculation. For specific fields, *T* can be set based on consensus by the peers.

Assigning of Threshold value (*T*) is the base for the evaluation technique. Any physicochemical experiments assume the temperature and/or pressure others at a fixed level. The can increase one or more parameters and result can be for every increase or decrease in temp or press. Also normal for some regions will be 20 degree Celsius for others, it may be 25 degree Celsius or more.

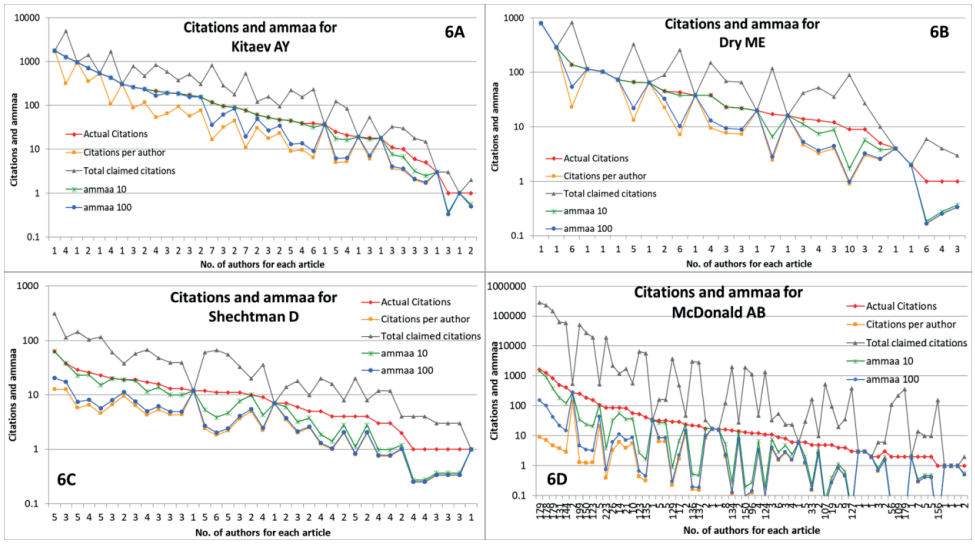


Figure 6. The relationship between number of coauthors for each article verses total citations, citations per author of the article, a total of the citations claimed by the coauthors, *ammaa 10* and *ammaa 100*.

Here, the evaluation is for every researcher. S/he may be Nobel Laureates or others who may have hundreds of publications or researchers/scholars/students who may have just fewer publications.

So, different *T* values have to be assigned for different set of researchers, institutions, etc. – *T* value can be anything. At the same time, all researchers may have to be evaluated for their impact. In such cases, a universal one is needed. Here, *T* it is assumed to be 100. The manuscript provides such provision which may be changed, now or anytime in future.

Conclusion

The new metrics treats all coauthors are equally in measuring the impact of the article. For every article of a researcher, *ammaa* can be calculated on the availability of a number of coauthors and number of citations. If an article authored by a single author, ammaa equals total citations of the article. For the coauthored articles, ammaa provides exponential credit from the first citation to the citations where (*aT-T*) intersect with the total citations. When number citations exceed (*aT-T*), all authors are credited with full citations. Threshold limit, *T* has to be fixed for universal standard and also on the case by case basis such as for the researchers of a country, region, organization and also for every subject. Further, *T* can be

increased or decreased depending on the circumstances. In the long term, the metrics may serve as a control in limiting the number of coauthors.

It can be applied to the past, present and future databases. Further, it can be used for citations, number of downloads, number of views, number of sharing, etc.

Acknowledgment
The author acknowledges the Director, CSIR Central Leather Research Institute for the permission to publish the article. The author also expresses his thanks to Saravanan for editing the text.

References

Ausloos, M. (2013). A scientometrics law about co-authors and their ranking: the co-author core. *Scientometrics, 95*(3), 895-909. doi:10.1007/s11192-012-0936-x

Cooper, C. (2003). Revisiting coauthor responsibility. *Science, 299*(5606), 511-511.

Cronin, B., Shaw, D., & La Barre, K. (2003). A cast of thousands: Coauthorship and subauthorship collaboration in the 20th century as manifested in the scholarly journal literature of psychology and philosophy. *Journal of the American Society for Information Science and Technology, 54*(9), 855-871. doi:10.1002/asi.10278

Cummings, J. N., & Kiesler, S. (2005). Collaborative research across disciplinary and organizational boundaries. *Social Studies of Science, 35*(5), 703-722. doi:10.1177/03061270505535

Egghe, L., Rousseau, R., & Van Hooydonk, G. (2000). Methods for accrediting publications to authors or countries: Consequences for evaluation studies. *Journal of the American Society for Information Science, 51*(2), 145-157. doi:10.1002/(sici)1097-4571(2000)51:2<145::aid-asi6>3.0.co;2-9

Figg, W. D., Dunn, L., Liewehr, D. J., Steinberg, S. M., Thurman, P. W., Barrett, J. C., & Birkinshaw, J. (2006). Scientific collaboration results in higher citation rates of published articles. *Pharmacotherapy, 26*(6), 759-767. doi:10.1592/phco.26.6.759

Galam, S. (2011). Tailor based allocations for multiple authorship: a fractional gh-index. *Scientometrics, 89*(1), 365-379. doi:10.1007/s11192-011-0447-1

Gazni, A., Sugimoto, C. R., & Didegah, F. (2012). Mapping World Scientific Collaboration: Authors, Institutions, and Countries. *Journal of the American Society for Information Science and Technology, 63*(2), 323-335. doi:10.1002/asi.21688

Hagen, N. T. (2010). Harmonic publication and citation counting: sharing authorship credit equitably - not equally, geometrically or arithmetically. *Scientometrics, 84*(3), 785-793. doi:10.1007/s11192-009-0129-4

Ilakovac, V., Fister, K., Marusic, M., & Marusic, A. (2007). Reliability of disclosure forms of authors' contributions. *Canadian Medical Association Journal, 176*(1), 41-46.

Iribarren-Maestro, I., Lascurain-Sanchez, M. L., & Sanz-Casado, E. (2007). Are multi-authorship and visibility related? Study of ten research areas at Carlos III University of Madrid. *Proceedings of ISSI 2007: 11th International Conference of the International Society for Scientometrics and Informetrics, Vols I and II*, 401-407.

Jack, G. (2015). Is mass authorship destroying the credibility of papers? Retrieved from https://www.timeshighereducation.com/news/mass-authorship-destroying-credibility-papers#node-comments

Leimu, R., & Koricheva, J. (2005). Does scientific collaboration increase the impact of ecological articles? *Bioscience, 55*(5), 438-443. doi:10.1641/0006-3568(2005)055[0438:dsciti]2.0.co;2

Liu, X. Z., & Fang, H. (2012). Fairly sharing the credit of multi-authored papers and its application in the modification of h-index and g-index. *Scientometrics, 91*(1), 37-49. doi:10.1007/s11192-011-0571-y

Porter, A. L., & Rafols, I. (2009). Is science becoming more interdisciplinary? Measuring and mapping six research fields over time. *Scientometrics, 81*(3), 719-745. doi:10.1007/s11192-008-2197-2

Roux, D. J., Rogers, K. H., Biggs, H. C., Ashton, P. J., & Sergeant, A. (2006). Bridging the science-management divide: Moving from unidirectional knowledge transfer to knowledge interfacing and sharing. *Ecology and Society, 11*(1), 20.

Sarna-Wojcicki, D., Perret, M., Eitzel, M., & Fortmann, L. (2017). Where Are the Missing Coauthors? Authorship Practices in Participatory Research. *Rural Sociology, 82*(4), 33. doi:https://doi.org/10.1111/ruso.12156

Seddon, P. J., Armstrong, D. P., & Maloney, R. F. (2007). Developing the science of reintroduction biology. *Conservation Biology, 21*(2), 303-312. doi:10.1111/j.1523-1739.2006.00627.x

Sekercioglu, C. H. (2008). Quantifying coauthor contributions. *Science, 322*(5900), 371-371.

Shaban, S. (2007). Multiple authorship trends in prestigious journals from 1950 to 2005. *Saudi Medical Journal, 28*(6), 927-932.

Shen, H. W., & Barabasi, A. L. (2014). Collective credit allocation in science. *Proceedings of the National Academy of Sciences of the United States of America, 111*(34), 12325-12330. doi:10.1073/pnas.1401992111

Tarnow, E. (2002). Coauthorship in physics. *Science and Engineering Ethics, 8*(2), 175-190. doi:10.1007/s11948-002-0017-2

Tol, R. S. J. (2011). Credit where credit's due: accounting for co-authorship in citation counts. *Scientometrics, 89*(1), 291-299. doi:10.1007/s11192-011-0451-5

Tscharntke, T., Hochberg, M. E., Rand, T. A., Resh, V. H., & Krauss, J. (2007). Author sequence and credit for contributions in multiauthored publications. *Plos Biology, 5*(1), 13-14. doi:10.1371/journal.pbio.0050018

Vieira, E. S., & Gomes, J. (2011). An impact indicator for researchers. *Scientometrics, 89*(2), 607-629. doi:10.1007/s11192-011-0464-0

Wager, E., Singhvi, S., & Kleinert, S. (2015). Too much of a good thing? An observational study of prolific authors. *Peerj, 3*. doi:10.7717/peerj.1154

Waltman, L. (2016). A review of the literature on citation impact indicators. *Journal of Informetrics, 10*(2), 365-391. doi:10.1016/j.joi.2016.02.007

Xu, J., Ding, Y., & Malic, V. (2015). Author Credit for Transdisciplinary Collaboration. *Plos One, 10*(9), 19. doi:10.1371/journal.pone.0137968

Yager, K. (2007). Each co-author should sign to reduce risk of fraud. *Nature, 450*(7170), 610-610. doi:10.1038/450610a

The impact of preprints in Library and Information Science: citations, usage, and social attention

Zhiqi Wang[1,2], Wolfgang Glänzel[3,4], Yue Chen[5]

[1]*zhiqi_wang90@126.com*
WISE Lab, Institute of Science of Science and S&T management, Dalian University of Technology, Dalian 116085 (China)

[2]*zhiqi.wang@student.kuleuven.be*
ECOOM, KU Leuven, Naamsestraat 61, Leuven, 3000 (Belgium)

[3]*Wolfgang.Glanzel@kuleuven.be*
ECOOM and Dept MSI, KU Leuven, Naamsestraat 61, Leuven, 3000 (Belgium)

[4]*glanzw@iif.hu*
Library of the Hungarian Academy of Sciences, Dept. Science Policy & Scientometrics, Budapest (Hungary)

[5]*chenyuedlut@163.com*
WISE Lab, Institute of Science of Science and S&T management, Dalian University of Technology, Dalian 116085 (China)

Abstract

The objective of this paper is to explore the impact of preprints in scholarly and broader scientific communication. In particular, the following four indicators are used to examine the 550 arXiv and 5782 non-arXiv papers in three major journals in Library & Information Science (LIS): citations from Web of Science Core Collection (WoS), Scopus and Google Scholar, usage counts in WoS, Mendeley readers and Tweets. The results show that arXiv papers have significant citation advantage across WoS, Scopus and Google Scholar in each year. Google Scholar provides statistically significantly larger number of citations and more 'early citations' than Scopus and WoS, but does not reflect greater citation advantage for arXiv papers. The impact advantage of arXiv papers can also be observed in Mendeley, but to a much lesser extent in WoS usage counts and in Tweets, indicating that arXiv papers gain broader attention than non-arXiv papers not only from users of the WoS. Mendeley readership as well as the usage counts in WoS have strong correlations with WoS citations, which are much stronger than those of Tweets. We can also conclude that unlike citations, information derived from statistics on users, readers and social media needs further exploration and in the case of social media also proper context analysis.

Introduction

Preprints gain increasing importance in scholarly communication, above all in fields like mathematics and physics, but preprints have recently become popular in the community of Library and Information Science (LIS) as well. However, the question to what extent preprints actually promote scholarly communication in various subject fields is still not fully addressed and analysed. The Open Access (OA) advantage on the citation impact of preprints has long been widely debated, and positive as well as negative results are found in several fields (Lawrence, 2001; Moed, 2007; Davis, 2011; Gargouri et al., 2010; Larivière et al., 2014), yet little is still known in LIS.

Citations from Clarivate Analytics Web of Science (WoS) are the most commonly used measures to assist research revaluation in terms of scholar impact at various levels of aggregation, such as institutions, journals or individuals (Glänzel et al., 2016; Zhou and Leydesdorff, 2011; Glänzel et al., 2014). Elsevier Scopus is also a

well-established alternative to the Web of Science, which has been used in Times Higher Education ranking (Harzing and Alakangas, 2016; SJR, 2018; National Science Board, 2018). Google Scholar is used as another alternative as it provides broader coverage of most discipline and higher index metrics than WoS and Scopus (Moed et al, 2016; Harzing and Alakangas, 2016), however, the reliability of its metrics and its applicability to higher levels of aggregation is still controversial (Thelwall and Kousha, 2017b).

A recent study (Chen et al. 2017) has found a WoS citation advantage of preprints in LIS, but a still necessary comprehensive investigation of the underlying evaluation results may be conducted by using different databases, especially for the preprints, which represent types of communication channels different from those indexed in traditional bibliographic databases. In this study, we will therefore compare the citations to arXiv papers and non-arXiv papers and their annual change between the two papers types and across the different data sources, particularly, WoS, Scopus and Google Scholar.

In addition to citation-based indicators, other metrics are used to extend measurement of scientific impact and to supplement traditional citation analysis (Sud and Thelwall, 2014), among these, *usage metrics* and *altmetrics* are the most common ones (Glänzel and Gorraiz, 2015). *Usage metrics* provide usage information of scientific articles, for instance, abstract or full-text views and downloads, which, in particular, provide a direct way to explore the usage preference of users. By using the download statistics from ScienceDirect regarding documents in Elsevier's electronic journal *Tetrahedeon Letters* and citations reported by WoS during a time interval of two years after publication, Moed (2005) found that one citation corresponded to about 100 downloads, and similar result was also found by Glänzel and Heeffer (2014) in *Physics A*. A further important observation concerns Open Access. Wang et al. found that OA papers enjoyed a more enduring attention from readers compared with the non-OA papers published in the same journal (Wang, et al., 2014; Wang et al., 2015). Thelwall and Kousha (2017a) analysed the view counts for articles uploaded to the academic social website ResearchGate and found that they have weak to moderate positive correlation with Scopus citation rates and Mendeley readership, however they did not identify whether the articles were uploaded to ResearchGate as full-text or not. Since September 2015, WoS provides the daily-updated usage counts of indexed documents on its platform to measure the level of interest in these documents. The "Usage Count" is defined as "the number of times the article has met a user's information needs as demonstrated by clicking links to a full-length article at the publisher's website (via direct link or OpenURL) or by saving the article for use in a bibliographic management tool (via direct export or in a format to be imported later)"[1] in the last 180 days or since 1 February 2013, providing a kind of brand new and all-round data source (Wang et al, 2016). Though several studies have studied the relationship between *Usage Count* and *Citation Impact* (Wang et al., 2016; Chi and Glänzel, 2017), the peculiarities of usage patterns for arXiv papers and non-arXiv papers indexed in WoS as reflected the by the user behaviour, has not yet sufficiently been studied. Addressing this question could help us better understand the role of different dissemination platforms in scholar communication. For papers both indexed in WoS and deposited in arXiv, we may proceed from initial assumption that the supplementary Open-Access accessibility provided by arXiv might reduce their usage

[1] https://images.webofknowledge.com/WOKRS529AR7/help/WOS/hp_usage_score.html

counts in WoS to a lower level than in the case of WoS indexed documents without pre-print versions.

The term of "altmetics" was introduced later than "usage metrics" (Priem et al., 2010), aiming at capturing new and previously invisible impacts of scholarly publications (Mohammadi and Thelwall, 2014). *Altmetrics* are metrics for measuring various activities on social web platforms, having the advantages of recognising a wider audience than just publishing scholars (Bornmann and Haunschild, 2015; Maflah and Thelwall, 2016). Among those, Mendeley readers and Twitter mentions are commonly analysed. Mendeley is a popular free global reference-manager tool and academic social network launched in 2009. The number of bookmarks of an article is regarded as Mendeley readership indicating reading, or the intention to read the articles that are bookmarked (Mohammadi et al. 2016). In this study, Mendeley readers are chosen as one of the indicators to reflect the impact of preprints, because, on one hand, many studies have verified that they show a strong positive correlation with citations to articles in various fields in the medical (Thelwall and Wilson, 2016) and nature sciences (Li, Thelwall & Giustini, 2012), as well as in the social sciences and humanities (Mohammadi and Thelwall, 2014). In addition, the coverage of Mendeley seems to be excellent for articles published in journals in information & library science (LIS) according to the WoS subject categories, as 58% LIS articles WoS-indexed articles in 2008 were also covered by Mendeley (Mohammadi and Thelwall, 2014), and the percentage was as high as 97.3% for JASIST articles published in the period 2001-2011 (Bar-Ilan, 2014). The other altmetric indicator, we intend to select, is Twitter mentions, which is more related to the impact in social media. A previous study substantiated that the extent and frequency of Twitter mentions is statistically correlated with arXiv downloads and 'early citations' in the initial period after the publication of a preprint (Shuai et al, 2012). Furthermore, comparing the total number of Twitter and Facebook posts of OA paper and non-OA papers in *Nature Communications*, Wang et al. (2015) concluded that OA articles attracted somewhat more social media attention than non-OA articles. The increasing share of preprints published in journals immediately suggests the necessity of further exploration of the impact of this publication model by altermetrics means to extend our notion of what is behind the mechanism of scholarly and wider impact.

This objective of this paper is therefore to study the impact of preprints in scholarly and wider scientific communication along the following research questions.

1. What is the (citation) impact differential for arXiv papers in WoS, Scopus and Google Scholar, and which one gives the most citation advantage for arXiv papers compared with non-arXiv papers?
2. Do arXiv papers take impact advantage from Usage according to WoS and altmetric attention, if so, to what extend as compared with citation advantage?
3. What is the relationship between the usage and altmeric-attention indicators with citations for arXiv and non-arXiv papers, and can certain characteristics be in the relationships for these two data sets be identified?

Data and Methods

For this study we use the same data sets as already used in the previous paper (Wang et al., 2018), where we had collected 550 arXiv papers and 5782 non-arXiv papers three the LIS journals *Scientometrics*, *Journal of the Association for Information Science and Technology* and *Journal of Informetrics* published in the period 2005–2017, but we excluded 288 Open Access articles in the three journals, which may

differ in terms of citations or other impact terms from non-OA papers and form a third category besides the toll access papers with and without preprint versions. The citations the 508 arXiv papers and 5536 non-arXiv papers have received are collected from the WoS and Scopus databases. Since Google Scholar did not permit automated data collection, the *Publish or Perish* (Harzing, 2007), which provides all necessary information at journal level, including DOI, titles, authors and citations, was used. Results were matched with the records in WoS based on the articles' DOIs. Articles that could not be matched by DOI (because of incorrect or missing identifier) were processed manually to extract citations from Google Scholar. The Mendeley reader counts and Tweets are collected from the open API. Table 1 gives the number of collected papers covered by the five data sources. All the indicators are cumulative, and the date when we collected data is 6th April, 2018. Papers that are not indexed in Scopus or Google Scholar, were excluded when the two sources were considered, and for those papers without Mendeley readership or Twitter posts *readers* and the *tweets* were put zero. The percentage of Mendeley readership for arXiv and non-arXiv papers published in the three journals indexed in WoS amounts to 97.05% and 96.66% respectively, the share of papers with Twitter posts, however, is with only 56.6% and 27.9% for arXiv and non-arXiv papers, respectively, much lower.

Table 1. The number of papers (2005-2017) covered by the five data sources (excluding OA papers)

	WoS	*Scopus*	*Google Scholar*	*Mendeley*	*Twitter*
arXiv	508	508	506	493	287
non-arXiv	5536	5482	5535	5351	1545

Methods and results

Citation advantage in Web of Science, Scopus and Google Scholar

Table 2 shows the citation advantage of arXiv papers obvious compared with non-arXiv papers across all three data sources (WoS, Scopus and Google Scholar) in annual breakdown. In order to quantify and measure the advantage we used the *Citation Impact Differential* (CID), which has been proposed by Moed (2005) and already applied to citation analysis of preprint publications by Wang et al. (2018). Its formula can be found in the legend of Table 1. In line with the usual trends, the total number of citations according to Google Scholar exceeds those found in Scopus and WoS, which are, otherwise, fairly similar. This is also shown by the semi-log graph in Figure 1 along with the decreasing citation rates as a consequence of the shortening citation windows with growing publication year. Figure 1 reveals an interesting trend, particularly, that the difference for citations in Google Scholar compared with citations in WoS and Scopus is more pronounced in the case of non-arXiv than arXiv papers. A possible reason may be that arXiv papers usually receive higher citation rates in WoS and Scopus than non-arXiv papers, while a similar strong effect could not be observed in Google Scholar, where information is mainly collected from repositories. Furthermore, the CID values of arXiv and non-arXiv papers derived from different data sources in Table 2 show that, although Google Scholar has a broader coverage than the other two databases and is based on links to the contents of digital repositories (Moed et al., 2016; Harzing and Alakangas, 2016), it does not provide larger citation advantage for arXiv papers. Nevertheless, Google Scholar still provides larger research metrics,

which may be of advantage, especially in an evaluative context at the individual-article level (Thelwall and Kousha, 2017b).

Table 2. The citation advantage of arXiv papers vs non-arXiv papers

PY	*WoS*			*Scopus*			*Google Scholar*		
	arXiv	*non-arXiv*	*CID**	*arXiv*	*non-arXiv*	*CID**	*arXiv*	*non-arXiv*	*CID**
2005	39.9	24.96	46.07	43.8	30.83	34.76	83.4	60.64	31.60
2006	123.6	29.26	123.43	139.8	35.63	118.76	271.9	77.77	111.04
2007	75.63	28.24	91.25	87.42	36.43	82.34	150.05	76.05	65.46
2008	49.89	22.82	74.46	54.47	26.81	68.06	91.42	54.94	49.85
2009	56.13	20.87	91.58	63.72	26.6	82.20	120.47	50.95	81.11
2010	50.78	20.56	84.72	58.22	25.47	78.27	113.54	45.15	86.19
2011	38.12	18.21	70.69	42.97	22.31	63.30	77.24	39.61	64.41
2012	31.66	14.14	76.51	35.13	17.06	69.25	60.72	31.03	64.72
2013	15.97	9.32	52.59	18.05	11.44	44.83	30.67	20.31	40.64
2014	17.16	7.68	76.33	20.48	9.22	75.82	39.71	17.1	79.60
2015	13.07	5.09	87.89	16.45	6.15	91.15	34.33	11.14	102.00
2016	7.95	3.07	88.57	9.77	3.98	84.22	18.22	8.7	70.73
2017	1.99	0.71	94.81	2.61	1.17	76.19	6.03	2.4	86.12

$^{*}CID = 100 \cdot (CPP_a - CPP_{na})/((CPP_a + CPP_{na})/2)$
CPP_a: *the number of received citations per paper in arXiv*
CPP_{na}: *the number of received citations per paper not in arXiv*

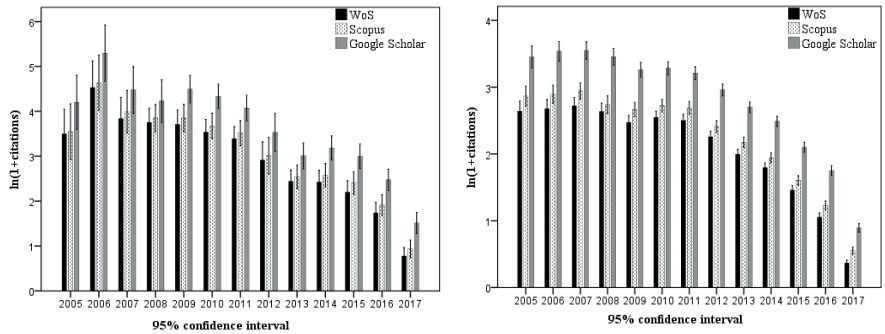


Figure 1. Semi-log presentation of citation counts and 95% confidence intervals for arXiv papers (left) and non-arXiv papers (right)

The citation advantage we applied in the previous part was designed to compare the impact between the two document "types", but we did not yet take another important aspect into consideration, namely the time dimension to capture the characteristics of citations to older and newer articles. The results are shown in Figure 2 as the trend lines based on a modified version of Citation Impact Differential, denoted by CID2, which has been proposed by Wang et al. (2018) to allow the comparison of the citation advantage across different data sources. Its formula can be found in the legend of Figure 2. Not surprisingly, the CID values for arXiv papers are lower than non-arXiv in consideration of the result presented in Figure 1, but it is interesting to see that the CID values of GS-WoS – "arXiv Google Scholar" and "non-arXiv Google Scholar"– start increasing considerably since 2013 and although the CID values of

SC-WoS show the similar trend as GS-WoS, they are not at the same level. Google scholar thus proved useful to find more 'early citations' than WoS and Scopus.

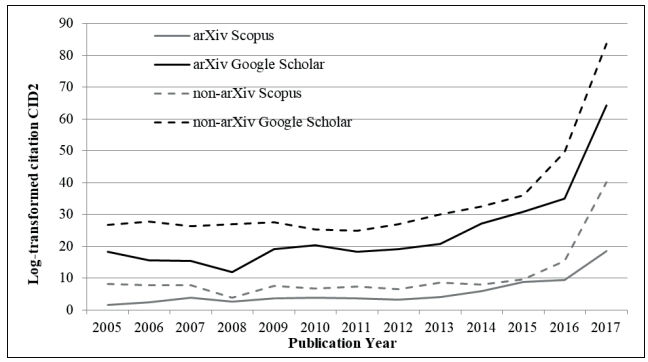


Figure 2. The log-transformed citation advantage for arXiv papers and non-arXiv papers

$CID2 = 100 * (CPP_{SC\backslash GS} - CPP_{WoS})/((CPP_{SC\backslash GS} + CPP_{WoS})/2)$

$CPP_{SC\backslash GS}$: $the\ average\ log - tranformed\ citations\ in\ Scopus(SC)/Google\ Scholar(GS)$

CPP_{WoS}: $the\ average\ log - tranformed\ citations\ in\ Web\ of\ Science(WoS)$

The usage and social attention advantage of arXiv papers

In this part, which represents a completely new approach to the analysis of the bibliometrics effects of preprint publishing, we explore the altmetric impact. In particular, we analyse the usage counts in WoS, Mendeley readership and Twitter posts, to address and answer our second research question. The mean values of Mendeley readers, Twitter mentions and usage counts in WoS are listed in Table 3, and the impact advantage is quantified by the *Altermetric Impact Differential* (AID), a version of the previous CID indicator adopted to the actual needs for the altemetrics. The formula is given in the legend of Table 3.

Table 3. The usage counts in WoS, Mendeley readers and Tweets advantage of arXiv papers

PY	*Mendeley*			*Twitter**			*Usage***		
	arXiv	*Non-arXiv*	*AID****	*arXiv*	*Non-arXiv*	*AID*	*arXiv*	*Non-arXiv*	*AID*
2005	42.20	33.18	23.93	0.10	n/a	n/a	n/a	n/a	n/a
2006	73.70	36.12	68.43	0.20	n/a	n/a	n/a	n/a	n/a
2007	93.37	38.32	83.60	0.63	n/a	n/a	n/a	n/a	n/a
2008	67.11	35.99	60.36	0.05	n/a	n/a	n/a	n/a	n/a
2009	74.19	38.23	63.97	0.47	n/a	n/a	n/a	n/a	n/a
2010	83.63	43.26	63.62	0.37	0.20	58.29	n/a	n/a	n/a
2011	56.64	37.98	39.44	0.33	0.31	5.86	n/a	n/a	n/a
2012	66.53	33.59	65.82	15.13	1.59	162.01	n/a	n/a	n/a
2013	43.07	30.25	34.96	2.50	3.20	-24.48	51.48	47.02	9.06
2014	64.70	27.80	79.77	7.63	1.76	124.97	59.40	42.30	33.63
2015	66.85	25.02	91.06	7.25	1.76	121.76	48.29	38.50	22.55
2016	41.12	21.64	62.09	5.80	3.40	52.18	35.95	29.47	19.82
2017	23.22	10.84	72.68	6.73	5.83	14.36	20.88	19.17	8.52

*There are no tweets for non-arXiv papers between 2005 and 2009.

**The usage counts in WoS are counted since 1 February 2013.

***AID (Altermetric Impact Differential):

$$AID = 100 * \left(CPP_{a(U\backslash M\backslash T)} - CPP_{na(U\backslash M\backslash T)}\right)/\left(\left(CPP_{a(U\backslash M\backslash T)} + CPP_{na(U\backslash M\backslash T)}\right)/2\right)$$
$$CPP_{U\backslash M\backslash T}: The\ mean\ usage\backslash Mendeley\ readers\backslash Tweets\ of\ per\ arXiv\ paper$$
$$CPP_{na(U\backslash M\backslash T)}: The\ mean\ usage\backslash Mendeley\ readers\backslash Tweets\ of\ per\ non-arXiv\ paper$$

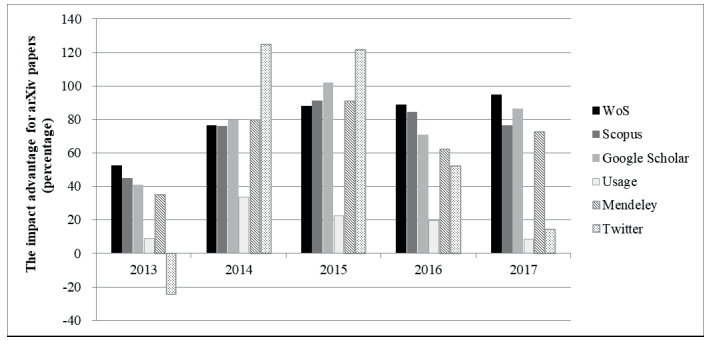


Figure 4. The impact advantage of arXiv papers

arXiv papers significantly have more Mendeley readers than non-arXiv papers, with AID values ranging from 23.9% to 91.1% in the period 2005–2017. By contrast, the AID values of usage counts for arXiv papers range between 8.5% and 33.6% representing a much lower level and indicating a much lesser advantage. The impact advantage from Tweets fluctuates in the wide range from -24.5% to 162.0%, which is most probably caused by the small underlying data set (cf. Table 1) and skewed and polarised distribution since only about half the arXiv-papers have tweets. This means, papers cannot really expect more tweets if authors deposit them on arXiv. The bars in Figure 4 present a comprehensive comparison of all the metric indicators we used in this paper, showing that the citation advantage of arXiv papers in LIS across all the three data sources is roughly the same, and it can be extended to Mendeley readers, but is hardly reflected by WoS usage counts and by Tweets, which actually proved very sparse.

The relationship between citations and usage in WoS and social attention

Regression analysis showed that the correlation of citation counts and the other impact measures may fluctuate over the years (cf. Thelwall and Wilson, 2016). Therefore and taking also the sample size of arXiv papers into consideration, we divided the entire period 2005–2017 into four sub-periods of different length and conducted regression analysis on the data subsets in the four sub-periods 2005–2009, 20010–2012, 2013–2015 and 2016–2017. The number of arXiv papers covered by WoS, Scopus and Google Scholar amounted to (91, 91, 90), (112, 112, 111), (178, 178, 177) and (127, 127, 127) respectively, and the corresponding number of non-arXiv papers ranges from 1092 to 1548. According to Table 4, the Mendeley readers as well as the usage counts in WoS are strongly correlated with citations in WoS for both data sets and all available periods, which is contrasted by the quite weak correlation of Tweets citation impact. The strongest correlation is the Mendeley readers and citations, indicating that the Mendeley readership has the most potential to be valuable as an import impact indicator for LIS articles with different publication age and publication model, either having a preprint version in arXiv or not. This result is very much in line with Thelwall's findings according to which Mendeley readership could serve as early impact indicator (Thelwall, 2018).

Table 4. Pearson correlation coefficients between citations in WoS//Scopus/Google and Mendeley readers, tweets and usage

PY	*WoS/Scopus/Google Scholar vs Mendeley*		*WoS/Scopus/Google Scholar vs Twitter**		*WoS/Scopus/Google Scholar vs Usage*	
	arXiv	*Non-arXiv*	*arXiv*	*Non-arXiv*	*arXiv*	*Non-arXiv*
2005-2009	.724**/.739**/.733**	.755**/.748**/.705**	0.06/0.07/0.08	n/a	n/a	n/a
2010-2012	.678**/.702**/.755**	.776**/.758**/.830**	0.10/0.10/0.10	.205**/.203**/.235**	n/a	n/a
2013-2015	.770**/.988**/.938**	.647**/.672**/.699**	.382**/.393**/.358**	.231**/.262**/.295**	.726**/.723**/.761**	.524**/.527**/.514**
2016-2017	.854**/.871**/.890**	.810**/.852**/.914**	.202*/.196**/.268**	.104**/.105**/.082**	.811**/825**/.843**	.622**/639**/.573**

The linear regression analysis in Figure 5 and 6, in which the four sup-periods are denoted by $*_1$ = 2005–2009 to $*_4$ = 2016–2017 (replace * by "a" for arXiv and "b" for non-arXiv papers), reveal another interesting aspect. The slopes, ranging between 0.1 and 0.75 in Figure 5, do not essentially differ between the two document "types". They can be considered kind of translation factor, i.e., the factor for ''translating'' Mendeley readership to citation impact (cf. Chi and Glänzel, 2018). The second observation in the context of the slope is that this factor decreases as time approaches present time, which is a consequence of the shrinking citation window, while the range of readership does by and large not change. By contrast, the slopes for usage vs. citations are different for the two data sets (see Figure 6), where the slopes for arXiv papers are about twice as those for non-arXiv papers. The specific translator factors for usage and citations in WoS have already been found and discussed by Chi and Glänzel (2018) and are also in line with our observations. The citations in WoS are shown as pars pro toto for citations because, on one hand, similar correlations can be found with citations from the other sources and, on the other hand, WoS covers the largest number of papers in our data set.

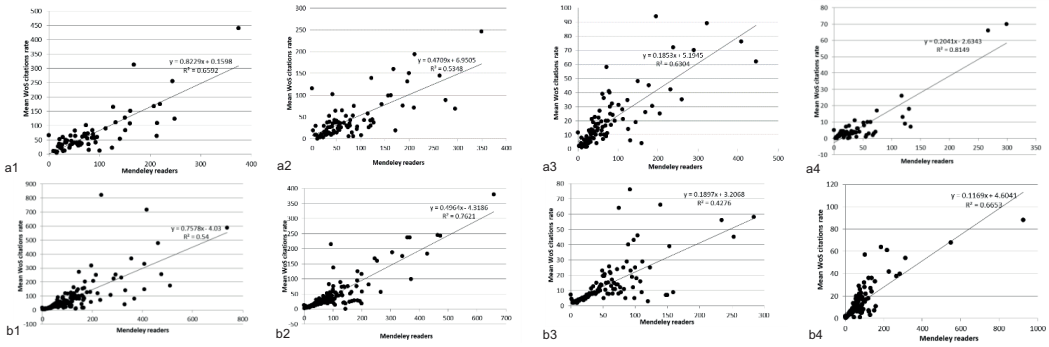
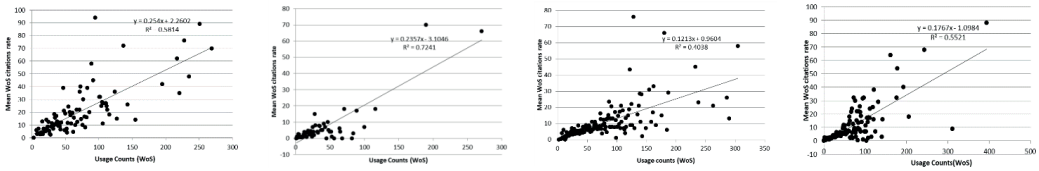


Figure 5. Plots of Mendeley readers vs citations in WoS in different time period for arXiv papers (a1-a4) and non-arXiv papers (b1-b4)



arXiv 2013-2015 **arXiv 2016-2017** **Non-arXiv 2013-2015** **Non-arXiv 2016-2017**

Figure 6. Plots of usage counts vs citations in WoS in different time period for arXiv papers and nonarXiv papers

In contrast to the annually reported citation counts in the WoS or Scopus databases, annual counts of altmeric data are not available. In order to have a proxy for the ageing of altmetrics when have to apply kind of a synchronous approach by calculating the means of indicator values per year and comparing the trends. The results are shown in Figure 7. There are significant distinctions between the trends of citations and social attention indicators. The series of green colour lines represent the arXiv papers and the orange colour lines represent the non-arXiv papers. In principle, trends for arXiv papers mirror those of non-arXiv papers but at a higher level. The deviation is the largest one in the case of Mendeley readership, most notably after 2013. Except for WoS citations, where the citation windows play an important role, all trends reflect growth. More recent article tend to attract more readers than older articles, as well as more usage counts from WoS users, which does not imply that older articles would not attract (new) readers anymore as can be seen from the first columns of Table 3. The magnitude of Twitter is much lower than readers, usage and even citations, however, with increasing trend. In this context we have to mention that both Usage and Readership does not reveal anything about the intention of the possible use of information, while the intention pursued by both citations and tweeds can be revealed from the textual context. As Glänzel and Chi (2019) concluded in their recent paper, the significance of social-media metrics do not yet meet the standards of traditional bibliometrics, and because of lacking reviewing and communication standards they require even more context analysis for meaningful interpretation than scholarly communication metrics.

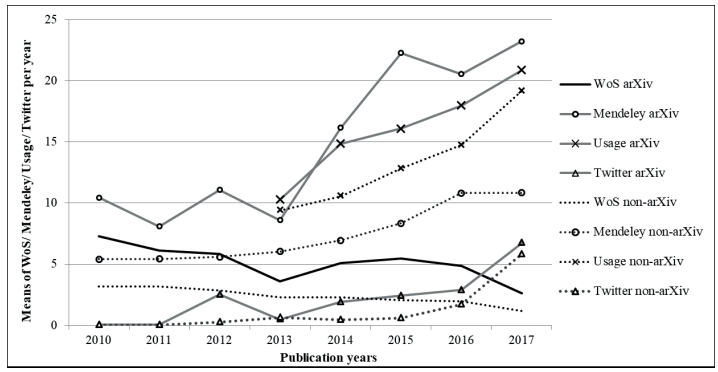


Figure 7. Mean values of WoS citations, Usage counts, Mendeley readers and Twitter posts to arXiv and non-arXiv papers (2010-2017)

Conclusions

The impact of preprints in LIS is studied by comparing citations, usage data and social-media attention between arXiv papers and non-arXiv papers. We have found significant citation advantage of arXiv papers across WoS, Scopus and Google Scholar in each year. Google Scholar provides statistically significantly larger number of citations than Scopus and WoS, but does not reflect greater citation advantage for arXiv papers. The longitudinal comparison of the Citation Impact Differential calculated between the two pars, Scopus-WoS and Google Scholar-WoS, suggests for both two data sets (arXiv and non-arXiv) that Google scholar provides more 'early citations' than WoS and Scopus, especially for non-arXiv papers.

For the usage and social attention metrics, we select the usage counts in WoS, Mendeley readers as well as Twitter mentions. The usage counts can be used as supplement or companion to the citation counts, most notably in the social sciences and humanities (cf. Chi and Glänzel, 2018; Chi et al., 2018). Bibliometric indicators based on Mendeley and Twitter are considered reflecting both academic and public interest in scientific literature. The impact advantage of arXiv papers can also be observed in Mendeley, but to a much lesser extent in WoS usage counts and in Tweets. For a longitudinal analysis we have subdivided the complete time period of nine years into four sub-periods with comparable sample sizes. A regression analysis has been conducted on the citation impact and the altmetric indicators. Similar patterns were found for both two data sets (arXiv and non-arXiv). Mendeley readership as well as the usage counts in WoS have strong correlations with WoS citations, which are much stronger than with Tweets. Another interesting observation was the "immediacy" of Mendeley readership, which confirms the already mentioned findings by Thelwall (2018). The "translation factor" for the different magnitudes of the variables expressed by the slope of the regression line confirms the observation by Chi and Glänzel (2018) in the context of usage and citations. Furthermore, the slope of the regression line of the WoS usage metrics for arXiv papers is two times as large as that for non-arXiv papers, and also greater than the Mendeley-WoS coefficient in the same period, indicating that arXiv paper gain broader attention than non-arXiv papers not only from users of the WoS. To answer our main question, there is evidence of impact advantage from preprint publishing in LIS not only in terms of scholarly communication but also in the mirror of altmetrics, except for Twitter posts. Unlike in the case of citations, where meaning and interpretation of metrics is sufficiently known, information derived from statistics on users, readers and social media needs further exploration and in the case of social media also proper context analysis. This will be part of future research.

References

Bornmann, L. & Haunschild, R. (2015). Which people use which scientific papers? An evaluation of data from F1000 and Mendeley. *Journal of Informetrics*, *9*(3), 477–487.

Chi, P. S. & Glänzel, W. (2017). An empirical investigation of the associations among usage, scientific collaboration and citation impact. *Scientometrics*, *112*(1), 403–412.

Chi, P.S. & Glänzel, W. (2018). Comparison of citation and usage indicators in research assessment in scientific disciplines and journals. Scientometrics 116(1), 537-554.

Chi, P. S., Gorraiz, J. & Glänzel, W. (2018). Comparisons of altmetric, usage and citation indicators in chemistry disciplines. . In *Proceedings of the 23rd*

International Conference on Science and Technology Indicators (pp. 174-183). CWTS, Leiden University.
Chen, Y., Wang, Z., Tan, J., Liu, Z.(2017). The position of preprint in scholarly communication: A biliometric and empirical study of arXiv. In *Proceedings of the ISSI conference 2017* (pp. 799–809).Wuhan University.
Davis, P. M. (2011). Open access, readership, citations: a randomized controlled trial of scientific journal publishing. *The FASEB Journal*, *25*(7), 2129–2134.
Gargouri, Y., Hajjem, C., Lariviére, V., Gingras, Y., Carr, L., Brody, T., & Harnad, S. (2010). Self-selected or mandated, open access increases citation impact for higher quality research. *PLoS ONE*, *5*(10), e13636.
Glänzel, W., Thijs, B. & Debackere, K. (2014). The application of citation-based performance classes to the disciplinary and multidisciplinary assessment in national comparison and institutional research assessment. *Scientometrics*, *101*(2), 939–952.
Glänzel, W. & Heeffer, S. (2014). Cross-national preferences and similarities in downloads and citations of scientific articles: A pilot study. In E. Noyons (Ed.), *Context counts: Pathways to master big and little data*. Proceedings of the STI conference 2014 (pp. 207–215). CWTS, Leiden University.
Glänzel, W. & Gorraiz, J. (2015). Usage metrics versus altmetrics: confusing terminology? *Scientometrics*, *102*(3), 2161–2164.
Glänzel, W., Thijs, B. & Chi, P. S. (2016). The challenges to expand bibliometric studies from periodical literature to monographic literature with a new data source: the book citation index. *Scientometrics*, *109*(3), 2165–2179.
Glänzel, W. & Chi, P. S., (2019). *Research beyond scholarly communication – the big challenge of Scientometrics 2.0*. Paper submitted for presentation at ISSI 2019, Rome, 2–5 September 2019.
Harzing, A.W. (2007). *Publish or Perish.* Retrieved December 28, 2018 from: https://harzing.com/resources/publish-or-perish.
Harzing, A. W. & Alakangas, S. (2016). Google Scholar, Scopus and the Web of Science: a longitudinal and cross-disciplinary comparison. *Scientometrics*, *106*(2), 787–804.
Larivière, V., Sugimoto, C. R., Macaluso, B., Milojevic, S., Cronin, B. & Thelwall, M. (2014). ArXiv e-prints and the journal of record: An analysis of roles and relationships. *Journal of the Association for Information Science and Technology*, *65*(6), 1157–1169.
Lawrence, S. (2001). Online or invisible. *Nature*, *411*(6837), 521. Retrieved from http://ivyspring.com/steveLawrence/SteveLawrence.htm
Li, X., Thelwall, M. & Giustini, D. (2012). Validating online reference managers for scholarly impact measurement. *Scientometrics*, *91*(2), 461–471.
Maflah, N. & Thelwall, M. (2016). When Are Readership Counts as Useful as Citation Counts? *Journal of the Journal of the Association for Information Science and Technology*, *67*(1), 191–199.
Moed, H. F. (2005). Statistical relationships between downloads and citations at the level of individual documents within a single journal. *Journal of the American Society for Information Science and Technology*, *56*(10), 1088–1097.
Moed, H. F. (2007). The effect of "open access" on citation impact: An analysis of ArXiv's condensed matter section. *Journal of the American Society for Information Science and Technology*, *58*(13), 2047–2054.
Moed, H. F., Bar-Ilan, J. & Halevi, G. (2016). A new methodology for comparing Google Scholar and Scopus. *Journal of Informetrics*, *10*(2), 533–551.

Mohammadi, E. & Thelwall, M. (2014). Mendeley Readership Altmetrics for the Social Sciences and Humanities: Research Evaluation and. *Journal of the Association for Information Science and Technology*, *65*(8), 1627–1638.

Mohammadi, E., Thelwall, M. & Kousha, K. (2016). Can Mendeley bookmarks reflect readership? A survey of user motivations. *Journal of the Association for Information Science and Technology*, *67*(5), 1198–1209.

National Science Board (2018), *Science & Engineering Indicators 2018.* Retrieved January 02, 2019 from https://www.nsf.gov/statistics/2018/nsb20181/report.

Priem, J., Taraborelli, D., Groth, P., Neylon, C. (2010), *Altmetrics: A manifesto*. Retrieved January 02, 2019 from http://altmetrics.org/manifesto.

Shuai, X., Pepe, A. & Bollen, J. (2012). How the Scientific Community Reacts to Newly Submitted Preprints: Article Downloads, Twitter Mentions, and Citations. *PLoS ONE*, *7*(11), 1–8.

SJR (2018), *Scimago Journal & Country Rank*. Retrieved January 02, 2019 from https://www.scimagojr.com/.

Sud, P. & Thelwall, M. (2014). Evaluating altmetrics. *Scientometrics*, *98*(2), 1131–1143.

Thelwall, M. & Kousha, K. (2017a). ResearchGate Articles: Age, Discipline, Audience Size, and Impact. *Journal of the Association for Information Science and Technology*, *68*(2), 468–479.

Thelwall, M. & Kousha, K. (2017b). ResearchGate versus Google Scholar: Which finds more early citations? *Scientometrics*, *112*(2), 1125–1131.

Thelwall, M. & Wilson, P. (2016). Mendeley Readership Altmetrics for Medical Articles: An Analysis of 45 Fields. *Journal of the Journal of the Association for Information Science and Technology*, *67*(8), 1962–1972.

Thelwall, M. (2018). Early Mendeley readers correlate with later citation counts. *Scientometrics*, 115(3), 1231-1240.

Wang, Z., Glänzel, W. & Chen, Y.(2018). How self-archiving influences the citation impact of a paper: A bibliometric analysis of arXiv papers and non-arXiv papers in the field of information science and library science. In *Proceedings of the 23rd International Conference on Science and Technology Indicators* (pp. 323-330). CWTS, Leiden University.

Wang, X., Fang, Z. & Sun, X. (2016). Usage patterns of scholarly articles on Web of Science: a study on Web of Science usage count. *Scientometrics*, *109*(2), 917–926.

Wang, X., Liu, C., Mao, W. & Fang, Z. (2015). The open access advantage considering citation, article usage and social media attention. *Scientometrics*, *103*(2), 555–564.

Wang, X., Mao, W., Xu, S. & Zhang, C. (2014). Usage history of scientific literature: Nature metrics and metrics of Nature publications. *Scientometrics*, *98*(3), 1923–1933.

Zhou, P. & Leydesdorff, L. (2011). Fractional counting of citations in research evaluation: A cross- and interdisciplinary assessment of the Tsinghua University in Beijing. *Journal of Informetrics*, *5*(3), 360–368.

Who acknowledges who? A gender analysis

Adèle Paul-Hus[1], Philippe Mongeon[2], Maxime Sainte-Marie[1] and Vincent Larivière[3]

[1] adele.paul-hus@umontreal.ca ; maxime.sainte-marie@umontreal.ca
École de bibliothéconomie et des sciences de l'information, Université de Montréal, PO Box 6128, Downtown Station, Montreal, Quebec, H3C 3J7 (Canada)

[2] philippe.mongeon@ps.au.dk
Danish Centre for Studies in Research and Research Policy, Department of Political Science, Aarhus University, Aarhus, Denmark

[3] vincent.lariviere@umontreal.ca
École de bibliothéconomie et des sciences de l'information, Université de Montréal, PO Box 6128, Downtown Station, Montreal, Quebec, H3C 3J7 (Canada) and Université du Québec à Montréal, Centre Interuniversitaire de Recherche sur la Science et la Technologie (CIRST), Observatoire des Sciences et des Technologies (OST), CP 8888, Succ. Centre-Ville, H3C 3P8 Montréal, Qc. (Canada)

Abstract

Acknowledgements found in scholarly papers allow for credit attribution among non-authors—from individuals to organizations—that contributed to a piece of research. As such, they are associated with a different kind of recognition than authorship. While several studies have shown that social factors affect authorship and citation practices, few analyses have been performed on acknowledgements. Based on 878,250 acknowledgees mentioned in 291,167 papers published in 2015 retrieved from Web of Science, the objective of this work-in-progress is to better understand such credit attribution practices and how gender may influence them. Our results show that gender disparities generally found in authorship can be extended to acknowledgements, and that women are even more underrepresented in the acknowledgement than in the authors' list of a paper. Our findings also confirm that women acknowledge proportionally more women. These results suggest that gender plays a role in the entire spectrum of credit attribution practices, from authorships to acknowledgements.

Introduction

Acknowledgements found in scientific papers are a public testimony of authors' gratitude and recognition that can help reveal contributions of varied nature made by individuals, institutions and organizations. The rich information acknowledgements convey allows credit distribution among authors and other contributors. However, in the reward system of science (Merton, 1973) where authorship constitutes the main means to accumulate "symbolic capital" (Bourdieu, 1975), a mention in the acknowledgement is not associated to the same kind of recognition as authorship. Given the hierarchical structure of the scientific community, credit attribution can be difficult to disentangle from one's status within the hierarchy (Heffner, 1979). Who will be named as an author and who will be in the acknowledgements of a publication? Credit attribution criteria do not only rely on the nature of the contribution made, and numerous other factors come into play, namely the disciplinary context and sociodemographic variables which have been shown to affect one position within the hierarchy, such as gender, age and academic status (Merton, 1973; Cole and Cole, 1973; Zuckerman, 1977).

Heffner (1979) was one of the few to investigate scientific credit attribution using acknowledgements. Based on a questionnaire completed by 207 acknowledgees from social and natural sciences, Heffner found that publication credit is not always accorded on the basis of the norm of universalism. More specifically, he showed that female PhDs were twice more likely than any other group (male and non-PhDs) to believe that they had been excluded from co-authorship when they felt their contribution warranted a place in the byline. Along those lines, Moore (1984) investigated the effect of authors' gender on the content of their acknowledgements and, more specifically, on the gender of those acknowledged. Using a sample of 300 male-authored and 70 female-authored psychology books, Moore found that

while men acknowledged mainly other men for their contribution and advice, women acknowledged the contributions of both genders. In another analysis, based on 684 psychology articles, Moore (1984) found a lower proportion of female acknowledgees, especially among articles from male authors. The author concludes that "there is a tendency on the part of each sex to seek out and acknowledge the professional advice of same-sex colleagues. For women, this most often means acknowledging individuals of both sexes (a practice relatively infrequent among male authors)" (p. 1029). More recently, Sugimoto and Cronin (2012) came to the same conclusion while analyzing the scholarly production of six important information scientists. They found that the authors included in their sample were more likely to acknowledge individuals of the same sex. Since those, few studies have focused on the issue of gender in credit attribution practices using acknowledgment data.

Objective and Research Questions

Focusing on the individuals named in acknowledgments, the objective of this work-in-progress is to better understand credit attribution practices of researchers and how gender may influence those. More specifically, we aim at answering the following research questions:

- What proportion of acknowledgees are women?
    - How does it vary as a function of the proportion of women in the byline of the acknowledging paper?
    - How does it vary as a function of the gender of the leading authors?
    - How does it vary by discipline?

Material and methods

Data

This study is based on all acknowledgements extracted from 2015 articles and reviews indexed in the Science Citation Index Expanded (SCI-E) and Social Sciences Citation Index (SSCI) from Clarivate Analytics' Web of Science (WoS). Access to the WoS data in a relational database format was provided by the *Observatoire des sciences et des technologies* (http://www.ost.uqam.ca). Acknowledgements data are collected and indexed by WoS only if they include funding information (Paul-Hus, Desrochers and Costas, 2016). These data are structured in three fields: the 'Funding Text' (FT), 'Funding Agency' (FO) and 'Grant Number' (FG). The dataset used in the present analysis was extracted from the FT field, which is the full text as it appears in the paper from which it is retrieved, and includes a total of 1,009,411 acknowledgements texts from as many papers. The dataset covers all disciplines included in the SCI-E and SSCI: Biology, Biomedical Research, Chemistry, Clinical Medicine, Earth and Space, Engineering and Technology, Health, Mathematics, Physics, Professional Fields, Psychology and Social Sciences. The discipline of a paper was assigned using the NSF field classification of journals (National Science Foundation, 2006); the NSF classification assigns only one discipline specialty to each journal, thus preventing potential multiple counting of papers.

Analysis

The extraction of individual names from acknowledgement texts was done using the Stanford Named Entity Recognizer (NER) (Finkel et al., 2005) module of the Natural Language ToolKit (NLTK) (Bird, 2009). To obtain the number of acknowledgee names per paper, the algorithm was applied on each string of acknowledgement text retrieved from the FT field and all named entities tagged as 'person' were selected. [1]

[1] This extraction procedure was used on a previous round of analysis of this dataset (Paul-Hus, Mongeon, Sainte-Marie & Larivière, 2017).

Several data cleaning procedures were then undertaken in order to eliminate non-human entities from the list of extracted names. First, incomplete names were removed from the list (entities containing only a first or last name, or only initials), retaining only entities composed of a complete name (i.e. full first and last name). To remove all remaining names that did not refer to individual persons such as grant, foundation, organization and institution names, manual cleaning was performed on the list. Examples of such names removed by manual cleaning include: Frederick Banting (grant), Marie Curie (grant and foundation), Boehringer Ingelheim (organization) and Instituto de Salud Carlos III (institution). Since acknowledgements often contain the name(s) of the author(s) signing the paper from which the acknowledgements are retrieved, a final step of cleaning was necessary. When the name(s) extracted from the acknowledgements of a paper X matched the name of one of the authors appearing in the byline of that paper X (using the first initial and the last name), this name was removed from the acknowledgees list for that specific paper, such as in the example below:

> Paper X
> Authors: J. Zhang, X. Feng and Y. Xu
> Acknowledgements text: "Jinsong Zhang, Xiao Feng, and Yong Xu contributed equally to this work […]."

For the purpose of our analysis, we consider first and corresponding authors as lead authors of a paper since first authors are often associated with the highest proportion of tasks performed in a paper (Larivière et al., 2016) and corresponding authors—who are in charge of correspondence—are often associated to conception and supervision (Mattsson et al., 2011). If both the first and corresponding authors are women, the paper is considered female-led, if both are men, the paper is considered male-led and if first and corresponding author are of different genders, the paper is considered mixed.

The gender assignation of personal names (authors and acknowledgees) was done using the Wiki-Gendersort algorithm (Bérubé, Ghiasi and Larivière, in preparation). Using Wikipedia pages to get gender information, this algorithm increases reliability of gender assignation by examining the first names of the names covered by Wikipedia and counting the number of masculine and feminine pronouns in the introduction section of the first twenty pages. Gender is assigned to the first name when the same gender was attributed to 75% of Wikipedia pages. No gender is assigned when this threshold is not met. Using the Wiki-Gendersort algorithm, we were able to identify the gender of 75% of personal names in our dataset (authors and acknowledgees). The final dataset includes 878,250 acknowledgees mentioned in 291,167 papers.

Results

Figure 1 presents the global share of women among all authors and acknowledgees by discipline and shows that the well-known gender gap found in authorship (Larivière et al. 2013; West et al., 2013) is also present in the acknowledgements. Women represent less than 50% of authors and of acknowledgees in all disciplines, with the only exception of Health where women account for 54% of authors and 55% of acknowledgees. Moreover, Health, Psychology, Clinical Medicine and Biomedical Research—domains where women have been traditionally more present (Witz, 1992)—are the only disciplines where female acknowledgees represent a greater proportion than female authors. In all other disciplines, women are not only under-represented in authorships, they are even less present when looking at the acknowledgements.

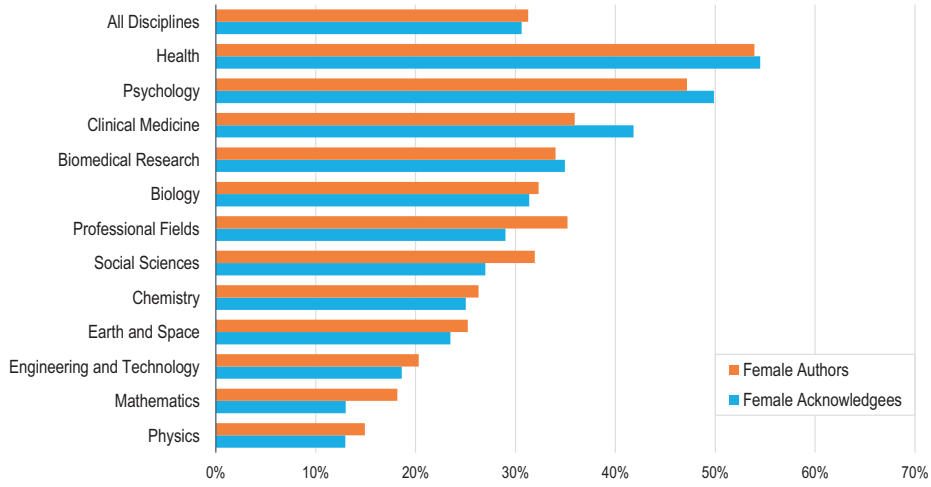


Figure 1. Proportion of female authors and acknowledgees, by discipline.

To further investigate the representation of women, we analysed the proportion of female acknowledgees per paper in relation to the proportion of women among authors of a paper. If the share of female acknowledgees is higher than the share of women among authors, female acknowledgees are considered over-represented in relation to authorship, and inversely. Figure 2 demonstrates that in most disciplines, women are more under-represented in the acknowledgements as compared to their proportional presence in authorships. On average, when all disciplines are considered, women are underrepresented by 3% as compared to their share of authorships. Clinical Medicine, Biomedical Research and Psychology are the only disciplines where the trend is reverse and where women are overrepresented in the acknowledgements as compared to their proportion of authorships. Altogether, Figure 2 shows that, in most disciplines, women are even more under-represented in the acknowledgements than they are in authorships.

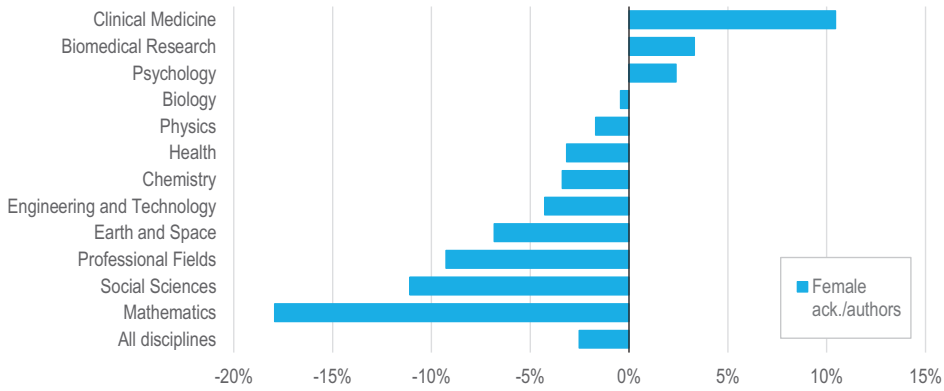


Figure 2. Proportion of female acknowledgees as a function of the proportion of female authors, by discipline.

Table 1 presents the proportion of acknowledgees who are women as a function of the gender of the leading authors. For all disciplines, the proportion of female acknowledgees is higher in female-led papers (women as first or corresponding authors) than in male-led papers or in the mixed group. The difference in the proportion of female acknowledgees between female and male-led papers ranges from 20.8% in the Professional Fields to 6.7% in Physics with a difference of 16.3% when all disciplines are taken together.

Table 1. Proportion of acknowledgees who are female as a function of the gender of the leading authors, by discipline.

Discipline	Female-led	Male-led	Mixed-led	Difference between female and male-led
Professional Fields	43.7%	22.9%	26.9%	20.8%
Health	61.3%	44.2%	52.3%	17.1%
Social Sciences	38.8%	22.0%	28.8%	16.7%
Clinical Medicine	52.1%	35.5%	43.7%	16.6%
Psychology	57.0%	41.4%	52.2%	15.6%
Chemistry	36.3%	21.8%	28.9%	14.5%
Mathematics	24.1%	10.5%	23.4%	13.7%
Engineering and Technology	28.1%	15.9%	26.5%	12.2%
Biomedical Research	43.1%	31.5%	37.9%	11.6%
Biology	37.3%	28.5%	35.3%	8.7%
Earth and Space	29.3%	21.6%	32.2%	7.7%
Physics	18.2%	11.5%	22.0%	6.7%
All Disciplines	42.0%	25.7%	36.9%	16.3%

Discussion and conclusion

Our results demonstrate that gender disparities generally found in authorship extend to acknowledgements. In most disciplines, women are proportionally more under-represented in the acknowledgement than in the authors' list of a paper. Furthermore, as previously found by Moore (1984) and Sugimoto and Cronin (2012), our findings clearly confirm that women acknowledge proportionally more women than men. These results seem to indicate that gender plays a role in credit attribution practices. In a broader context, the under-representation of women in the acknowledgements of scientific papers adds up to the global gender disparities found in authorships, citations and self-citations (Ghiasi, Mongeon, Sugimoto and Larivière, 2018; Larivière et al., 2013; West et al., 2013).

To further our understanding of credit attribution practices, the next steps of this analysis will focus on the relationship between acknowledgements and academic status, looking more specifically at the number of papers published by acknowledgees and the number of citations received as a proxy for the reputation of a researcher.

References

Bérubé, N., Ghiasi, G., & Larivière V. (in preparation). Assigning gender to first names using Wikipedia.

Bird, S., Klein, E., & Loper, E. (2009). *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*. O'Reilly Media, Inc.

Bourdieu, P. (1975). The specificity of the scientific field and the social conditions of the progress of reason. *Social Science Information, 14*(6), 19‑47.
Cole, J. R., & Cole, S. (1973). *Social stratification in science*. Chicago, IL: University of Chicago Press.
Finkel, J. R., Grenager, T., & Manning, C. (2005). Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics* (p. 363–370). Stroudsburg, PA, USA: Association for Computational Linguistics. https://doi.org/10.3115/1219840.1219885
Ghiasi, G., Mongeon, P, Sugimoto, C.R., & Larivière, V. (2018). Gender homophily in citations. In *Proceedings of the 23rd International Conference on Science and Technology Indicators (p.1520-1525).* Leiden, NL: Centre for Science and Technology Studies (CWTS). http://hdl.handle.net/1887/65291
Heffner, A. G. (1979). Authorship Recognition of Subordinates in Collaborative Research. *Social Studies of Science*, 9(3), 377 384. https://doi.org/10.1177/030631277900900305
Hoder-Salmon, M. (1978). Collecting Scholar's Wives. *Feminist Studies*, 4(3), 107 114. https://doi.org/10.2307/3177543
Mattsson, P., Sundberg, C.J., and Laget, P. (2011). Is correspondence reflected in the author position? A bibliometric study of the relation between corresponding author and byline position. *Scientometrics*, 87, 99–105.
Merton, R. K. (1973). *The sociology of science: Theoretical and empirical investigations*. Chicago and London: Chicago University Press.
Moore, M. (1984). Sex and Acknowledgements: A Nonreactive Study. *Sex Roles, 10* (11-12), 1021 1031.
National Science Foundation. (2006). *Science and Engineering Indicators*. Chapter 5: Academic Research and Development. Data and Terminology. http://www.nsf.gov/statistics/seind06/c5/c5s3.htm#sb1
Larivière, V., Ni, C., Gingras, Y., Cronin, B. & Sugimoto, C.R. (2013). Bibliometrics: Global gender disparities in science. *Nature*, 504(7479), 211-213
Larivière, V., Desrochers, N., Macaluso, B., Mongeon, P., Paul-Hus, A., and Sugimoto, C.R. (2016). Contributorship and division of labor in knowledge production. *Social Studies of Sciences*, 46, 417–435.
Paul-Hus, A., Mongeon, P., Sainte-Marie, M., & Larivière, V. (2017). The sum of it all: Revealing collaboration patterns by combining authorship and acknowledgements. *Journal of Informetrics*, *11*(1), 80-87. https://doi.org/10.1016/j.joi.2016.11.005
Paul-Hus, A., Desrochers, N., & Costas, R. (2016). Characterization, description, and considerations for the use of funding acknowledgement data in Web of Science. *Scientometrics*, 108(1), 167 182. https://doi.org/10.1007/s11192-016-1953-y
Sugimoto, C. R., & Cronin, B. (2012). Biobibliometric profiling: An examination of multifaceted approaches to scholarship. *Journal of the American Society for Information Science and Technology*, *63*(3), 450 468. https://doi.org/10.1002/asi.21695
West, J.D., Jacquet, J., King, M.M., Correll, S.J., Bergstrom, C.T. (2013). The role of gender in scholarly authorship. *PLoS ONE* 8(7): e66212. https://doi.org/10.1371/journal.pone.0066212
Witz, A. (1992). Professions and Patriarchy: Routledge.
Zuckerman, H. A. (1977). *Scientific elite: Nobel laureates in the United States*. Piscataway, N.J.: Transaction Publishers.

Accuracy of Policy Document Mentions: the Role of Altmetrics Databases

Francis Houqiang Yu[1] Cathy Xueting Cao[2] Tingting Xiao[3] and Zhenyi Yang[4]

[1] *yuhouq@yeah.net* [2] *caoxueting0829@163.com* [4] *zhenyi_yang@foxmail.com*
School of Economics & Management, Nanjing University of Science & Technology, Nanjing (China)

[3]*xtthahaha@gmail.com*
Jiangsu Institute of Quality and Standardization, Nanjing (China)

Abstract

Policy document is recognized as a promising source for indicating societal impact of scientific product. As data of high quality are the basis of reliable altmetrics research and applications, the accuracy of policy document altmetrics data remains to be critical yet unexplored problem. The study, based on in-depth coding analysis of 2079 records of policy document altmetrics data that are sampled from 79 policy document source platforms in Altmetric.com database, has revealed that 8% of records have errors produced by the author of policy document (type one error), 70% of records have errors produced by altmetrics database (type two error). For type two error, policy document link error (5%), fake policy document mention (13%), duplicate or omission of authors (38%) and publication date error (8%) are identified as the most severe errors. Underlying reasons of these errors and strategies of avoiding them are discussed.

Introduction

Altmetrics has offered non-traditional way of measuring the diverse impact of scientific product. Altmetrics indicators have been widely adopted by various stakeholders for showcasing the general attention and the potential impact of their work. Many previous studies focused on altmetrics data sources like Twitter (Mohammadi, Thelwall & Kwasny et al., 2018; Yu, 2017), Mendeley (Aduku, Thelwall & Kousha, 2017; Zahedi & Haustein, 2018) and F1000 (Bornmann & Haunschild, 2015). Lately, policy document is studied as a potential useful data source for indicating the societal impact of scientific product (Bornmann, Haunschild & Marx, 2016), because policy usage of scientific product is supposed to reflect the relationship between academic research and policy making. The policy document is defined very broadly by the Altmetric.com company and refers to any policy, guidance, or guidelines document from a governmental or non-governmental organization[1]. Compared with other altmetrics data sources, policy document is particularly of interest in twofold, i.e. it is target-oriented and it focuses on a relevant part of society. Wooldridge and King (2018) found that altmetric score is highly correlated with the peer review scores of societal impact and news and policy sources are likely to contain the most informative data to assess impact.

Despite the popularity of altmetrics, many fundamental questions are not answered among which data quality is identified as one of the major challenges (Haustein, 2016). According to NISO (2016), accuracy, transparency and replicability are three essential dimensions of data quality. Altmetrics databases are the infrastructure of altmetrics research and services. They constantly collect, clean and store altmetrics data in large scale and provide the data in a systematic and usable way. Altmetric.com database is by far the most commonly used altmetrics database.

Data quality of citation databases

Good quality of bibliographic data is fundamental to bibliometrics research and

[1] See https://help.altmetric.com/support/solutions/articles/6000060968-what-outputs-and-sources-does-altmetric-track-

application. Abundant studies have been dedicated to examine the data quality of citation databases. These studies may provide relevant implication for studies of altmetrics data quality. To sum up, there are two perspectives of studying the topic.

The first perspective is to directly check the accuracy of bibliographic data in the database via content analysis. Investigated databases are mainly multidisciplinary databases like Web of Science (WoS), Scopus and Google Scholar (GS). Data errors can be classified into two major categories (Buchanan, 2006), i.e. author error and database mapping error. Author error refers to error that is generated by the cited author, for example, the error can occur in the publication title, publication year, volume number or pagination. Many studies have investigated the first category of data error and reported its percentage. The percentage varies according to the subject area and the scope of author error as defined by the researcher. Based on a large-scale study, Moed (2002) estimated the percentage of author error in SCI database to be 7%. Database mapping error refers to error that is generated by citation databases, for example, transcription error, or cited-article omitted from a cited-article list. However, only a few studies have looked at the second category of data error, because in the past it is difficult to determine the rate of clerical errors introduced in the citation database (Garfield, 1974). Franceschini, Maisano and Mastrogiacomo (2014) analyzed the omitted citations and found the omitted-citation rate of Web of Science to be 6%. Many weird errors are discovered in Scopus (Franceschini, Maisano & Mastrogiacomo, 2016a). The database mapping error is further classified into errors in the transcription of author names or article title, incomplete cited-article list, omitted cited-article list, wrong or missing DOI, and errors concerning online-first papers (Franceschini, Maisano & Mastrogiacomo, 2016b).

The second perspective is to compare different bibliographic databases. Meho and Yang (2007) compared the data quality of three citation databases, i.e. WoS, Scopus and Google Scholar (GS) and found GS has low accuracy and thus low reliability. Calver et al. (2017) studied the discrepancy of searching result between Scopus, Google Scholar, WoS and WoSCC, and found that retrieval result using the same keyword is different for different databases. Therefore, to improve the consistency across different databases will help users reduce the cost of finding proper database and improve the retrieval efficiency. Doner (2017) studied the influence of inaccurate document type on the citation index and found that data of accurate and inaccurate document type would lead to different citation counts. In addition, several comprehensive studies (Archambault, Campbell & Gingras, 2009; Harzing & Alakangas, 2016; Wang & Waltman, 2016) had tapped into the data quality of Scopus, GS and WoS.

Meanwhile, the underlying reasons of the data error and corresponding improving strategies are discussed. It is suggested that authors should be more rigorous and self-disciplined in citing behavior for reducing data error (Zhao, 2009). Also, editors should put more efforts to check the authenticity and validity of references (Chen, 2014). In database level, dictionary database structure can be constructed and improved to automatically check the data with computers to reduce human cost (Su, 2001). Although not all discovered errors are corrected (Franceschini, Maisano & Mastrogiacomo, 2016c), bibliographic databases like Web of Science has made use of these studies to improve their data quality (Prins, Costas & van Leeuwen et al., 2016).

Data quality of altmetrics databases

Data quality is of critical importance for establishing credibility of altmetrics indicators. Without careful examination of the data quality, the application of altmetrics databases will be severely criticized. The topic is highlighted in altmetrics conference. Accuracy is above all the basis of data reliability.

Research of altmetrics data quality is still in preliminary stage. Most previous studies compare across different altmetrics aggregators to check the coverage of different altmetrics

data sources as collected and maintained by these aggregators (Chamberlain, 2013; Peters, Jobmann & Eppelin, 2014; Zahedi, Fenner & Costas, 2014). It is found that factors such as timing, platform and publication identifier will influence the reported altmetrics data (Meschede & Siebenlist, 2018). Even for the same set of papers, different major aggregators provide different metrics (Ortega 2018). The difference challenges the reliability of altmetrics. Zahedi and Costas (2018) investigated the underlying reasons and found that a range of different methodological, technical and reporting choices have determined the final counts, yet no universe recommendations could be made for data aggregators and users because each practice has its own cons and pros.

As regards the direct checking of the accuracy or completeness of altmetrics data, Zahedi, Haustein and Bowman (2014) used a small sample to compare the metadata of publications presented in Mendeley and indexed in Web of Science and found that journal and article titles are the most erroneous. Unlike citations, Mendeley reader count is fluctuate and may decrease (Bar-Ilan, 2014).

Studies of policy document altmetrics

Policy document usage of academic research is strong evidence of societal impact. According to (Newson, King & Rychetnik, 2018), two parallel streams of research are contributing to provide insights about the influence of research on policy, i.e. research impact assessments that starts from research, and research use assessments that starts from policy document. Commonly used data sources are interviews, surveys, policy documents, focus groups and direct observation. Case study is the most popular method of the policy document impact research. Many policy assessment studies use forward tracing approaches to examine single policy document to corroborate claimed impacts. However, these qualitative studies are limited by the data scale and high cost.

Policy document altmetrics has provided the opportunity to measure the usage of scientific product in policy document in a more systematic way. Several studies have explored policy document altmetrics. Bornmann and Haunschild (2016) find that only 1.2% of climate change publications have at least one policy mention. In another study, Haunschild and Bornmann (2017) find that less than 0.5% of papers indexed in Web of Science are mentioned at least once in policy-related documents. The percentage is lower than expected, because Khazragui and Hudson (2015) argue that policy tends to be based upon a large body of work. Nevertheless, Tattersall and Carroll (2018) used policy document altmetrics data collected by Altmetric.com company to evaluate the research from University of Sheffield. It is found that Altmetric.com company has offered important and highly accessible data on the policy impact, but the data must be used with caution because data errors are found.

Research questions

Although data quality of citation databases is well studied, data quality of altmetrics database remains to be explored. While policy document altmetrics can indicate the societal impact of scientific product, the accuracy of policy document altmetrics data is still unclear. The study is aimed to measure the accuracy of policy document altmetrics data in Altmetric.com database, in order to inform the potential limitation of research or application that are based on Altmetric.com policy document data and provide reference for improving the data quality in the future. To be specific, the questions are:

(1) Is there error in policy document altmetrics data? If so, what types of errors may occur?

(2) What is the distribution of different categories of errors in policy document altmetrics data?

The result is compared with the accuracy situation of traditional citation databases. Potential underlying reasons of the data error are discussed.

Methodology

Data source

As one of the major altmetrics data aggregators, Altmetric.com company has started to collect policy document mentions of scientific product in January 2013[2]. Data were retrieved on June 8th, 2018. In total, there are 1,390,350 policy document mentions of 1,072,397 unique scientific product, from 80,476 unique policy documents that are released by 79 unique source platforms. The top 10 source platforms are listed in Table 1.The length of policy document and the number of scientific products mentioned by the policy document are quite different. The longest policy document (entitled with *Dietary Reference Intakes: The Essential Guide to Nutrient Requirements*) has 1345 pages and mentioned 3731 unique scientific products.

Table 1 Productivity distribution of policy document source platforms (Top 10)

R.	*Policy document source platform*	*N.P.*	*%*	*N.M.*	*%*
1	World Health Organization	10935	13.6%	240140	19.1%
2	Analysis & Policy Observatory (APO)	9592	11.9%	123493	9.8%
3	National Academies Press	7565	9.4%	295981	23.5%
4	UK Government (GOV.UK)	7025	8.7%	57792	4.6%
5	Centers for Disease Control and Prevention (CDC)	6122	7.6%	157443	12.5%
6	National Bureau of Economic Research	4992	6.2%	46987	3.7%
7	World Bank	4132	5.1%	31316	2.5%
8	European Food Safety Authority	3405	4.2%	20829	1.7%
9	Food and Agriculture Organization of the United Nations	2813	3.5%	35467	2.8%
10	rijksoverheid.nl	2794	3.5%	25530	2.0%

*N.P. is number of policy documents of each source platform that are detected by altmetrics database, N.M is number of policy document mentions of scientific product.

Sampling strategy

To measure the accuracy of policy document mentions, we begin by analyzing the production process of altmetrics data, as is illustrated in Figure 1. Compared with traditional bibliometrics data, the production of altmetrics data has two additional steps, i.e. production of altmetrics source data (like policy document) by the altmetrics source platform (like policy document website), and production of the final altmetrics data in altmetrics database (like Altmetric Explorer) accomplished by altmetrics companies (like Altmetric.com company). Because the quality issue of bibliometrics data has been well studied, in this study, we focus only on these two additional steps. Both steps could potentially introduce data errors. The error caused by the altmetrics source platform is defined as type one error, and the error caused by the altmetrics company is defined as type two error.

[2] See https://help.altmetric.com/support/solutions/articles/6000136884-when-did-altmetric-start-tracking-attention-to-each-attention-source-

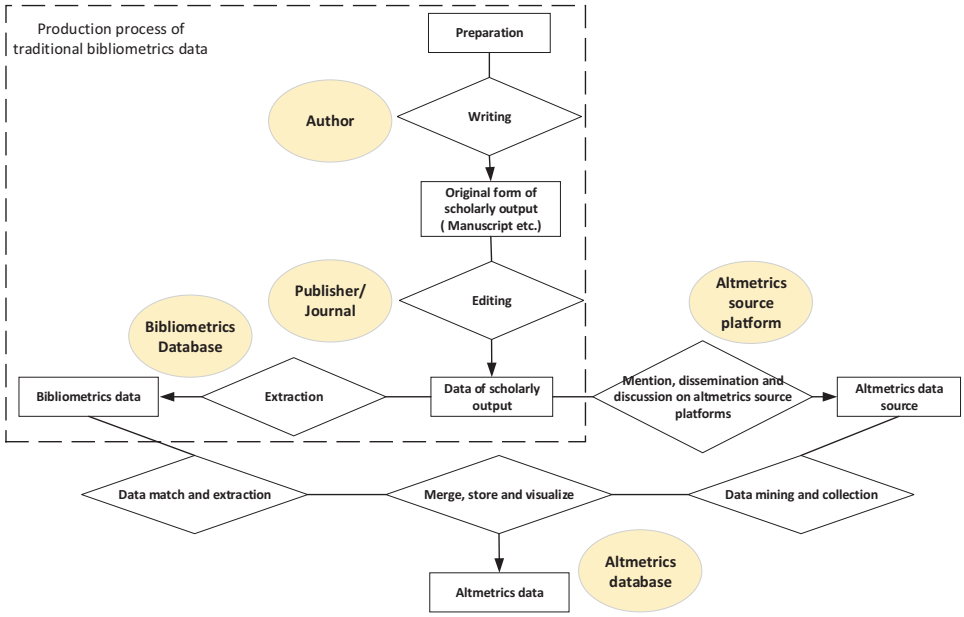


Figure 1 Production process of altmetrics data

According to Altmetric.com company, each policy document source platform has its unique data structure and therefore they have created a personalized crawler for each source platform. Except for that, all records are supposed to share the same data extraction and matching technique that the company has used to automatically collect, merge and visualize altmetrics data. In total, there are 79 unique policy document source platforms. 30 policy documents from each source platform are randomly selected. For policy document source platform of which the total number of captured policy documents is no higher than 30, all policy documents are selected. For each policy document, one record of policy document mention is randomly selected. Each record has the relevant data of a policy document mentioning a scientific product. The final sample is a dataset of 2079 records.

Process of coding a single record

The process of coding a single record is demonstrated in Figure 2. In general, there are three steps as described below.

(1) Click the URL of policy document that is collected and recorded by the Altmetric.com database. If the URL doesn't work, the title of the policy document is used to search for the policy document, in order to examine the existence of such policy document.

(2) After we find the original file of the policy document (mostly in pdf format), we download and open the file to search for the mentioned scientific product as recorded by the Altmetric.com database, in order to examine whether the scientific product is indeed mentioned in the policy document.

(3) After we find the location of scientific product (mostly papers) mentioned in the policy document, we cross check the bibliographic information of the scientific product from three sources, i.e. the policy document file, the Altmetric.com detail page and the original file of scientific product. The original file of scientific product is retrieved from full-text databases or open access repositories, and is used as benchmark with which any inconsistency is deemed as error. The error is further classified into different categories. For a certain number of records, two categories of errors could simultaneously occur and are coded. Very few records have more

than two categories of errors, only two categories of errors that are supposed to be more important are coded.

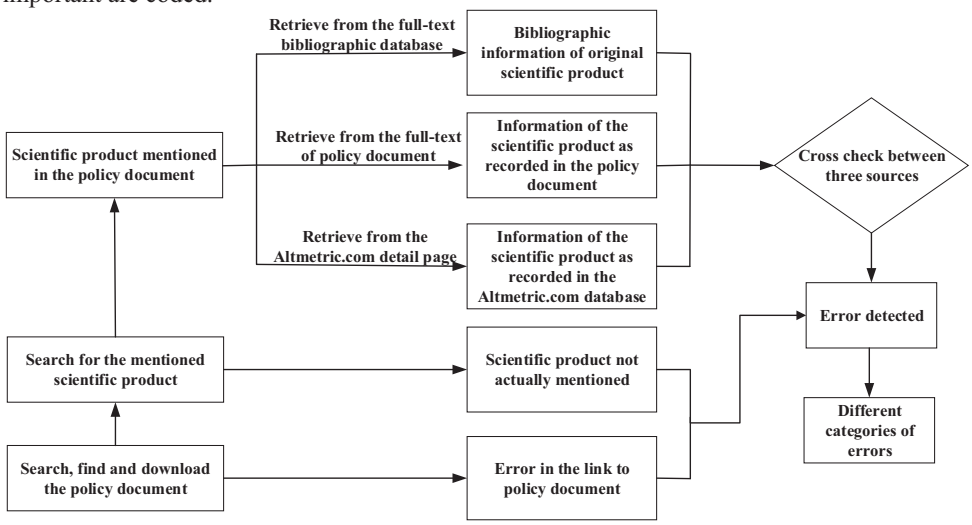


Figure 2 Process of examining a record of policy document

Process of coding the whole sample dataset

The coding table is established via three steps.

(1) 100 records are randomly selected from the sample dataset and are coded by four coders independently. The aim is twofold, to define the possible categories of errors and to standardize the coding process. Based on the coding practice, the initial coding table is created.

(2) Another 100 records are randomly selected from the sample dataset and are coded by the four coders using the initial coding table and the standardized coding process. The aim is also twofold, to improve the usability of the coding table and to calculate the coding consistency. A few new categories of error are added to the initial coding table. The rate of coding consistency between the four coders is 86%. The records of coding inconsistency are discussed to understand the underlying reason. Based on the discussion and new discovery of categories of error, the final coding table is obtained, as shown in Table 2 and Table 3.

Type one error is discovered by comparing the original file of scientific product and the information recorded in the policy document. Type two error is discovered by comparing the original file of scientific product, the policy document file and the Altmetric.com detail page.

Table 2 Coding table for type one error (errors caused by policy document source platforms)

Code	*Definition*
1.1 Author error	Author of scientific product wrongly recorded in the policy document
1.2 Title error	Title of scientific product wrongly recorded in the policy document
1.3 Publication date error	Publication date of scientific product wrongly recorded in the policy document
1.4 Source (Journal) error	Source (Journal)of scientific product wrongly recorded in the policy document

Table 3 Coding table for type two error (errors caused by Altmetrics database)

Code	*Definition*
2.1 Policy document link error	
2.1.1 Link updated or expired	Link to the policy document is updated or expired
2.1.2 No page found	Link to the policy document returns no page found
2.1.3 Link to page of multiple document	Link to the policy document leads to multiple versions of files
2.2 Fake policy document mention	
2.2.1 Coincidence of same text	Title of the scientific product coincides with text in the policy document
2.2.2 Policy document mentioned by itself	Mentioned scientific product is the recommended citing format of the policy document itself.
2.3 Transcription error	
2.3.1 Title error of scientific product	Title error of scientific product in the database
2.3.2 Author error of scientific product	
2.3.2.1 Omission	Author's name omitted
2.3.2.2 Misspell	Author's name misspelled
2.3.2.3 Duplicate	Author's name duplicated
2.3.3 Source (Journal) error of scientific product	Source (Journal) title error of scientific product in the database
2.3.4 Publication date error of scientific product	
2.3.4.1 Omission of publication date	Publication date omitted in the database
2.3.4.2 Inconsistency of publication date	Publication date inconsistent with the formal publication date
2.3.5 Title error of policy document	Title error of policy document in the database

(3) the total sample dataset of 2079 records are coded by two coders using the final coding table. The rate of coding consistency between the two coders is 98%. Records of inconsistent coding are discussed and coded with the help of a third coder.

Result

Distribution of data errors in the first level code

Distribution of data errors in the first level code is shown in Table 4. For the investigated dataset, 73% of records have data errors. The percentage is substantially high. 8% of records have type one error while 70% of records have type two error. 4% of records have both type one error and type two error.

Table 4 Distribution of data error in the first level code

First level code	*Number of incorrect records*	*%*
1 Type one error	169	8.1%
2 Type two error	1445	69.5%
Total	1523	73.2%

*Total number of records is 2079.

Distribution of data errors in the second level code

Table 5 shows the distribution of data errors in the second level code. For type one error,

code 1.1 error has the highest percentage (5%). It means that in every twenty policy document mentions of scientific product, one mention will wrongly record the author name of the mentioned scientific product.

For type two error, the transcription error is highlighted and takes up 52% of total records. It is the major reason of the high percentage of type two error. 13% of policy document mentions are fake because scientific product that is supposed to be mentioned is not mentioned at all. The situation of policy document link error is also poor, 5% of the links are no longer accessible.

Table 5 Distribution of data error in the second level code

Second level code	*N.*	*% of total incorrect records*	*% of the upper level code*
1.1 Author error of scientific product	107	5.2%	63.3%
1.2 Title error of scientific product	31	1.5%	18.4%
1.3 Publication date error of scientific product	19	0.9%	11.2%
1.4 Journal error of scientific product	12	0.6%	7.1%
2.1 Policy document link error	104	5.0%	7.2%
2.2 Fake policy document mention	261	12.6%	18.1%
2.3 Transcription error	1080	52.0%	74.8%

Distribution of data errors in the third level code

To take a closer look at type two error, Table 6 shows the distribution of data errors in the third level code. For code 2.1 policy document link error, majority of them is due to *no page found.* For code 2.2 fake policy document mention, it is mainly due to the coincidence of same text. For code 2.3 transcription error, most of the errors occur in author of the scientific product. To be specific, 38% of total records have error introduced by altmetrics database to the author of mentioned scientific product.

Table 6 Distribution of data error of type two error in the third level code

Third level code	*N.*	*% of total incorrect records*	*% of the upper level code*
2.1.1 Link updated or expired	5	0.2%	5%
2.1.2 No page found	92	4.4%	88%
2.1.3 Link to page of multiple document	7	0.3%	7%
2.2.1 Coincidence of same text	204	9.8%	78%
2.2.2 Policy document mentioned by itself	57	2.7%	22%
2.3.1 Title error of scientific product	71	3.4%	7%
2.3.2 Author error of scientific product	781	37.6%	72%
2.3.3 Source (Journal) error of scientific product	45	2.2%	4%
2.3.4 Publication date error of scientific product	161	7.7%	15%
2.3.5 Title error of policy document	22	1.1%	2%

Moreover, we zoom in to see the detail of error code 2.3.2 and 2.3.4, the result is shown in Table 7 and Table 8. It is found that code 2.3.2.3 (duplicate of author name) is the major reason of code 2.3.2 (author error of scientific product in altmetrics database). Meanwhile, code 2.3.4.2 (inconsistency of publication date) contributes the most to code 2.3.4 (publication date error of scientific product in altmetrics database).

Table 7 Distribution of data errors of error code 2.3.2

Fourth level code	*N.*	*% of total incorrect records*	*% of the upper level code*
2.3.2.1 Omission of author of scientific product	79	3.8%	10%
2.3.2.2 Misspell of author of scientific product	21	1.0%	3%
2.3.2.3 Duplicate of author of scientific product	681	32.8%	87%

Table 8 Distribution of data errors of error code 2.3.4

Fourth level code	*N.*	*% of total incorrect records*	*% of the upper level code*
2.3.4.1 Omission of publication date	21	1.0%	13%
2.3.4.2 Inconsistency of publication date	139	6.7%	87%

Discussion and conclusion

The policy document altmetrics data are analogous to citation data in that authors of policy document may produce errors in data of mentioned scientific product, just like authors of citing publication may produce errors in data of cited publication. Altmetrics databases use program to crawl and collect data of policy document, just like citation databases use program to identify and extract citation data from publications. The analogous relationship is illustrated in Figure 3.

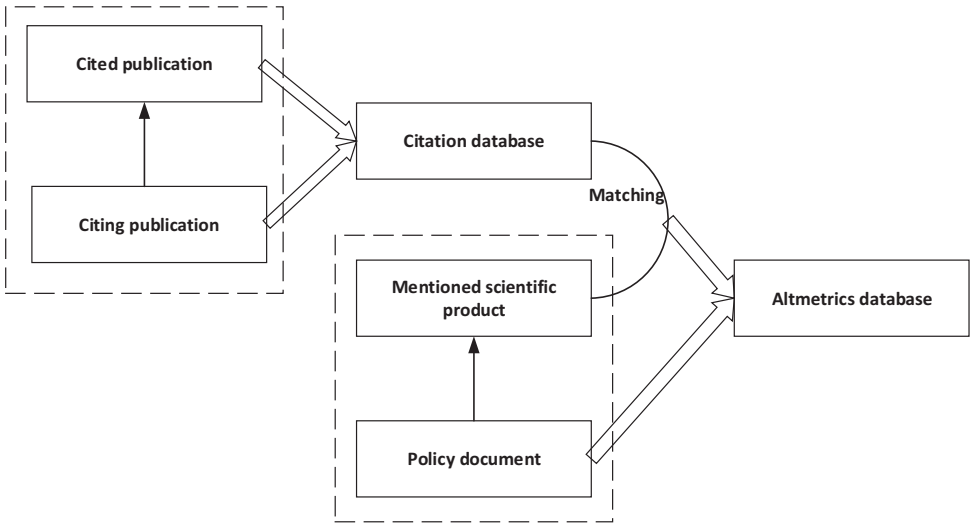


Figure 3 The analogous relationship between policy document altmetrics data and citation data

However, the difference between these two types of data is that altmetrics database extract data of scientific product from other databases rather than directly from the policy document. Therefore, only errors produced by the author of policy document and citing publication are comparable. Result is shown in Table 9. It is found that percentage of data errors of mentioned scientific product in policy document are all slightly higher than that of cited publications in citing publications.

Table 9 Comparison of errors introduced by authors of policy document and citing publications

Error by author of policy document		*Error by author of citing publication (Buchanan 2006)*	
Author error of mentioned scientific product	5.2%	Author error of cited publication	3.7%
Title error of mentioned scientific product	1.5%	Title error of cited publication	1.0%
Publication date error of mentioned scientific product	0.9%	Publication date error of cited publication	0.6%

Altmetrics databases rely on external bibliographic database to hold bibliographic data. In case of Altmetric.com database, CrossRef is the major bibliographic data provider. By fetching data from external bibliographic database, Altmetrics databases are able to avoid some errors produced by author of policy document. However, 70% of policy document records in Altmetric.com database have data error, i.e. type two error. It is much higher than traditional database mapping error of which the percentage is 3.5% (Franceschini, Maisano & Mastrogiacomo, 2016b).

Policy document link error (percentage is 5%) has corroborated that altmetrics data are highly influenced by the stability of source web platform. Once the platform has removed the link or retracted the policy document, the policy altmetrics data are gone forever.

Some policy document altmetrics data are misleading because scientific product is in fact not mentioned at all (percentage is 13%). To our observation, this usually (78% of cases) occurs when the title of scientific product is short and in plain words, and the full text of policy document happen to include text of such title but not refer to the scientific product. The underlying reason could be that for these records only title is used for matching purpose without further cross-check via other metadata. The false policy document mention occasionally (22% of cases) occurs when the policy document is simultaneously a scientific product and has contained its own recommended citation format. For example, *EFSA journal* publishes the scientific outputs of the European Food Safety Authority which is one of the major source of policy document. The recommended citation data in the policy document is mistakenly regarded as one policy document mention of itself! The underlying reason could be that Altmetric.com company has not recognized this problem. It can be avoided by excluding self-mentions.

Transcription error is the most commonly found errors (percentage is 52%). Transcription error in title of scientific product (percentage is 3%) is usually found to be omitting part of the original title, in many cases the part after colon, in a few cases merely the latter part of a long title. Transcription error in author of scientific product (percentage is 38%) is mostly due to duplicate of authors (87% of cases), followed by omission of authors (10% of cases). Only a few misspell of author's name occur (3% of cases) and is usually due to the non-standard character set or abbreviation inconsistency. The underlying reason could be program exceptions. Transcription error in publication date of scientific product (percentage is 8%) is more complex in that a scientific product could have multiple publication date (Haustein, Bowman & Costas, 2015). In this study, publication date of journal issue is used as standard publication date, hence many records in altmetrics database that use first online mention date as publication date are deemed erroneous. However, there are also a few cases where publication date data are missing. Transcription errors in journal of scientific product or title of policy document are less severe, also in forms of omission or inconsistency.

In practical usage of altmetrics data, type one error is not a problem because it is not part of altmetrics data and some of these errors are corrected in altmetrics database. For type two error, it seems many errors, such as frequently found duplicate or omission of authors, omission

of title, publication date, are easy to be avoided as long as the processing programs are improved. Other less frequently found errors are relatively more difficult to detect and the underlying reason is unknown. For policy document link error, it seems that the altmetrics database shall either remove related records that become inaccessible, or consider documenting policy document file for future use. To avoid fake policy document mentions, the altmetrics database shall pay additional attention to scientific product of short and plain title and cross-check the mention with other metadata, they may as well exclude self-mentions when a policy document is also a scientific product.

Results of the study have made it clear that altmetrics data, especially policy document altmetrics data, are far from being perfect. The error rate is much higher than that of traditional citation database. Future studies based on the data need to be aware of this limitation. And altmetrics database need to improve the data processing technique. Finally, due to high labor cost of manual coding work, the study is based on analysis of a relatively small sample of the policy document altmetrics data. This needs to be taken into consideration in interpreting the results of the study.

Acknowledgement

The authors would like to thank Mr. Longfei Li and Ms. Zihan Yin for helping conduct the coding work and Altmetric.com company for providing the data. The research is supported by National Natural Science Foundation of China (NO.71804067), Humanity and Social Science Foundation of Ministry of Education of China (18YJC870023) and the Fundamental Research Funds for the Central Universities (No.30918013107).

References

Aduku, K. J., Thelwall, M. & Kousha, K. (2017). Do Mendeley reader counts reflect the scholarly impact of conference papers? An investigation of computer science and engineering. *Scientometrics,* 112(1): 573-581.

Archambault, É., Campbell, D., Gingras, Y. & Larivière, V. (2009). Comparing bibliometric statistics obtained from the Web of Science and Scopus. *Journal of the American society for information science and technology,* 60(7): 1320-1326.

Bar-Ilan, J. (2014). JASIST@ Mendeley revisited. altmetrics14: expanding impacts and metrics, workshop at web science conference 2014.

Bornmann, L. & Haunschild R. (2015). Which people use which scientific papers? An evaluation of data from F1000 and Mendeley. *Journal of Informetrics,* 9(3): 477-487.

Bornmann, L., Haunschild, R. & Marx, W. (2016). Policy documents as sources for measuring societal impact: how often is climate change research mentioned in policy-related documents? *Scientometrics*, 109(3): 1477-1495.

Buchanan, R. A. (2006). Accuracy of cited references: The role of citation databases. *College & Research Libraries*, 67(4): 292-303.

Calver, M. C., Goldman, B., Hutchings, P.A., & Kingsford, R.T. (2017). Why discrepancies in searching the conservation biology literature matter. *Biological Conservation*, 213: 19-26.

Chamberlain, S. (2013). Consuming article-level metrics: Observations and lessons. *Information Standards Quarterly*, 25(2): 4-13.

Donner, P. (2017). Document type assignment accuracy in the journal citation index data of Web of Science. *Scientometrics*, 113(1): 219-236.

Franceschini, F., Maisano, D., & Mastrogiacomo, L. (2014). Scientific journal publishers and omitted citations in bibliometric databases: Any relationship? *Journal of Informetrics*, 8(3): 751-765.

Franceschini, F., Maisano, D., & Mastrogiacomo, L. (2016a). The museum of errors/horrors in Scopus. *Journal of Informetrics*, 10(1): 174-182.

Franceschini, F., Maisano, D., & Mastrogiacomo, L. (2016b). Empirical analysis and classification of database errors in Scopus and Web of Science. *Journal of Informetrics*, 10(4): 933-953.

Franceschini, F., Maisano, D., & Mastrogiacomo, L. (2016c). Do Scopus and WoS correct old omitted citations? *Scientometrics*, 107(2): 321-335.

Garfield, E. (1974). Errors-theirs, ours and yours. in Essays of an Information Scientist (Philadelphia: ISI Pr., 1977), 2: 80–81. Originally published in Current Contents (June 19, 1974): 5–6.

Harzing, A.-W. & Alakangas S. (2016). Google Scholar, Scopus and the Web of Science: a longitudinal and cross-disciplinary comparison. *Scientometrics,* 106(2): 787-804.
Haunschild, R. & Bornmann L. (2017). How many scientific papers are mentioned in policy-related documents? An empirical investigation using Web of Science and Altmetric data. *Scientometrics*, 110(3): 1209-1216.
Haustein, S. (2016). Grand challenges in altmetrics: heterogeneity, data quality and dependencies. *Scientometrics,* 108(1): 413-423.
Haustein, S., Bowman, T.D. & Costas, R. (2015). When is an article actually published? An analysis of online availability, publication, and indexation dates. Retrieved from https://arxiv.org/ftp/arxiv/papers/1505/1505.00796.pdf
Khazragui, H. & Hudson J. (2015). Measuring the benefits of university research: impact and the REF in the UK. *Research Evaluation*, 24(1): 51-62.
Meho, L. I. & Yang K. (2007). Impact of data sources on citation counts and rankings of LIS faculty: Web of Science versus Scopus and Google Scholar. *Journal of the American society for information science and technology*, 58(13): 2105-2125.
Meschede, C. & Siebenlist T. (2018). Cross-metric compatability and inconsistencies of altmetrics. *Scientometrics*, 115(1): 283-297.
Moed, H. F. (2002). The impact-factors debate: the ISI's uses and limits. *Nature,* 415(6873): 731.
Mohammadi, E., Kwasny, M. & Holmes, K. L.. (2018). Academic information on Twitter: A user survey. *PLoS ONE*, 13(5): e0197265.
Newson, R., King, L., & Rychetnik, L. et al. (2018). Looking both ways: a review of methods for assessing research impacts on policy and the policy utilisation of research. *Health research policy and systems*, 16(1): 54.
NISO. (2016). Outputs of the NISO Alternative Assessment Project. Retrieved from https://groups.niso.org/apps/group_public/download.php/17091/NISO+RP-25-2016+Outputs+of+the+NISO+Alternative+Assessment+Project.pdf
Ortega, J. L. (2018). Reliability and accuracy of altmetric providers: a comparison among Altmetric. com, PlumX and Crossref Event Data. *Scientometrics,* 116(3): 2123-2138.
Peters, I., Jobmann, A. & Eppelin, A. al. (2014). Altmetrics for large, multidisciplinary research groups: A case study of the Leibniz Association. *Libraries in the Digital Age (LIDA) Proceedings 13.*
Prins, A. A., Costas, R., van Leeuwen, T. & Wouters, P.F. (2016). Using Google Scholar in research evaluation of humanities and social science programs: A comparison with Web of Science data. *Research Evaluation*, 25(3): 264-270.
Tattersall, A. & Carroll C. (2018). What Can Altmetric.com Tell Us About Policy Citations of Research? An Analysis of Altmetric.com Data for Research Articles from the University of Sheffield. *Frontiers in Research Metrics and Analytics*, 2(9).
Wang, Q. & Waltman L. (2016). Large-scale analysis of the accuracy of the journal classification systems of Web of Science and Scopus. *Journal of Informetrics*, 10(2): 347-364.
Wooldridge, J. & King M. B. (2018). Altmetric scores: An early indicator of research impact. *Journal of the Association for Information Science and Technology*. https://doi.org/10.1002/asi.24122
Yu, H. (2017). Context of altmetrics data matters: an investigation of count type and user category. Scientometrics, 111(1), 267-283.
Zahedi, Z. & Costas R. (2018). General discussion of data quality challenges in social media metrics: Extensive comparison of four major altmetric data aggregators. *PLoS ONE*, 13(5): e0197326.
Zahedi, Z., Fenner, M. & Costas, R. (2014). How consistent are altmetrics providers? Study of 1000 PLOS ONE publications using the PLOS ALM, Mendeley and Altmetric. com APIs. altmetrics 14. Workshop at the Web Science Conference, Bloomington, USA.
Zahedi, Z. & Haustein S. (2018). On the relationships between bibliographic characteristics of scientific documents and citation and Mendeley readership counts: A large-scale analysis of Web of Science publications. *Journal of Informetrics*, 12(1): 191-202.
Zahedi, Z., Haustein, S. & Bowman, T. (2014). Exploring data quality and retrieval strategies for Mendeley reader counts. SIG/MET workshop, *ASIS&T 2014 annual meeting*, Seattle. Retrieved from: www.asis.org/SIG/SIGMET/data/uploads/sigmet2014/zahedi.pdf.
Chen X.J. (2014). Study of quality of references and its auditing methods. *Chinese Journal of Scientific and Technical Periodicals*, 25(9): 1145-1148. In Chinese.
Su X.N. (2001). Quality control of data in citation indexes. *Journal of Library Science in China*, 27(2): 76-78. In Chinese.
Zhao Q.M. (2009). Analysis of errors in refereces of scientific journals and the prevention strategies. *Editorial Friends*, (6): 47-49. In Chinese.

University research diversification effect on its citation-based performance: A study of Australian universities

Alireza Abbasi[1] and Hamid R. Jamali[2]

[1] *a.abbasi@unsw.edu.au*
The University of New South Wales (UNSW), Canberra (Australia)

[2] *h.jamali@gmail.com*
School of Information Studies, Charles Sturt University, Wagga Wagga (Australia)

Abstract

This paper analyses the relationship between Universities' research diversification breadth and depth and their citation-based performance and also ranking. Universities' diversification is measured based on the disciplinary ratings of universities in Excellence in Research for Australia (ERA) in 2015. The results showed a significant positive relationship between both university diversification metrics developed in this study and most of the university citation-based performance metrics obtained from InCites and QS ranking values. In other words, universities which are active in more disciplines and are rated high in terms of their research activity in those disciplines are more likely to have better citation impact performance and ranked higher.

Introduction

University ranking, as Lopez-Illescas, de Moya-Anegon and Moed (2011) argued, has gone through considerable development over the last decade so much that it would develop into a discipline. University ranking already has its own principles, organization of experts (IREG, the International Ranking Experts Group), conferences, workshops and publications. Besides national university rankings such as ERA in Australia, there are several international ranking systems such as Academic Rankings of World Universities (Liu and Cheng, 2005), QS World University Rankings (Times QS, 2009), SCImago Institutions Rankings (SCImago, 2010), Leiden World Ranking (CWTS, 2009), and Ranking Web of World Universities (CL-CSIC, 2009). All these rankings vary in terms of the type of data and indicators they use. However, a major trend in most of them is to consider ranking by fields. Universities put a great deal of effort into improving their positions in all of these evaluation systems.

The consideration of research fields in evaluations and rankings is vital as there has been criticism in the past about ignoring the disciplinary factors. For instance, Abbott and Doucouliagos (2004) analysed the research output generated by the Australian university research centres in economics and concluded that it is very important to take the specialization into account. Taking specialization/ concentration and diversification of research in universities into account is also very important for making financial planning (Filippini and Lepori, 2007) and for cost efficiency analysis (Johnes and Johnes, 2009; Thanassoulis et al., 2011). More recent studies such as the one by Lopez-Illescas, de Moya-Anegon and Moed (2011) emphasised that university specialization should be taken into account in their evaluation and rankings. More recently Robinson-Garcia and Jiménez-Contreras (2016) have criticised university rankings for suppressing disciplinary differences as reducing universities' research activities to a single dimension leads to poor judgment, although many rankings now present ranking by field.

Although there seems to be no doubt that disciplinary differences and specialization should be incorporated for university ranking and other evaluation purposes, we do not know enough about the effect of diversification or concentration of research on the performance of

universities and their positions in evaluations or rankings. An old study (Pianta and Archibugi, 1991) used the number of publications and citations to study how the size of national research in advanced countries is related to their degree of specialization by fields of science. They found a negative correlation between the two variables, with Japan and, to a lesser extent Italy, showing a specialization degree higher than expected. The results showed that countries such as the US, the UK, the Netherlands, and Switzerland that are well-established in terms of science had a lower than expected specialization degrees, which implied they had larger diversification. Over time, however, most countries reduced their scientific specialization, despite the recent research on patents and technological specialization (Pianta and Archibugi, 1991).

Moed et al. (2011) studied the relationships between university research performance and concentration of university research at an institutional level and at a national level. Their results showed that a larger publication output both at a national and an institutional level was associated with a higher performance measured by citation counts. They considered the number of publications as a measure of concentration. Their results indicated that concentration and performance were positively related in university research, but no evidence was found that more concentration of research among a country's universities or among an institution's main fields is associated with better overall performance. Their study revealed a "tendency that the research in a particular subject field conducted in universities specializing in other fields outperforms the work in that field in institutions specializing in that field. This outcome may reflect that it is multi-disciplinary research that is the most promising and visible at the international research front and that this type of research tends to develop better in universities specializing in a particular domain and expanding their capabilities in that domain towards other fields" (p. 649). In another study, Daraio, Bonaccorsi and Simar (2015) analysed data on input and output of some European universities and found out that economies of scale (i.e. size of universities) and specialization (i.e. covering many fields) both have a significant impact on the efficiency of the Humboldt model that includes a coexistence of teaching and research and coexistence of many disciplines within the same institution. Nevertheless, confirming previous findings, they found that specialization did not have a significant impact on the efficiency of the research model. In a more recent study of EU countries, Pastor and Serrano (2016) concluded that differences in field of science specialisation might be one of the influencing factors in considerable differences in output per capita (citable documents per R&D researcher) among the higher education institutions of EU countries.

Lopez-Illescas, de Moya-Anegon and Moed (2011) used a measure of disciplinary specialization (called Gini Index) for 50 Spanish universities, and showed the extent to which research papers are evenly distributed among disciplines in general universities while are more concentrated in particular disciplines in more specialised universities such as medical, agricultural or (poli-) technical focused universities. The outcome of their research was that categorization of universities according to their disciplinary specialization into three categories—general, moderately specialized and highly specialized—is useful and that disciplinary specialization should be taken into account in university rankings. Another study on European universities in the field of medical sciences (Bonaccorsi and Secondi, 2017) concluded that size had a negative impact on research quality, quality was higher in generalist universities and those with more international collaboration. Wolszczak-Derlacz (2017) studied 500 European and American universities in terms of efficiency and found that older European universities were more efficient, government funding and lower efficiency were correlated in Europe and that universities in wealthier regions of Europe and the U.S. were more efficient.

While specialization should be considered in the evaluations and rankings, an important question for universities might be whether they should go for larger diversity in their research

or they should specialize and concentrate their research and try to only excel in a few fields. This research aims to answer the following research question: 'how does diversification of research relate to citation-based performance and university ranking?' To do so, we will use citation impact indicators and data from Australian national evaluation system (ERA: Excellence in Research for Australia) and InCites database.

Methods

Excellence in Research for Australia (ERA) is the Australian national research evaluation system that rates universities at a disciplinary level based on their research performance on a five-point scale (1 to 5). ERA is done every three years and it takes into account research quality (citation analysis and/or peer review), research activity (output, income etc.), research application (commercialization, patents etc.). ERA uses the Australian and New Zealand's Fields of Research (FoR) that is increasingly being used in scientometric studies (Rousseau, 2018) and by citation databases (e.g. Dimensions). The rating is not based on organizational units (i.e. departments or schools) but based on the assignment of research to FoR codes by researchers. Each article is assigned to one FoR code or a combination of FoR codes (with a total of 100% assignment for each article) by researchers themselves. For instance, a university might not have a Library and Information Science department or school but if there are researchers in that university who do research that fall in this area (0807 FoR code) and if they assign their research output and activity to this code, the university will be assessed in this code. FoR has 22 two-digit codes, which are divided into 157 four-digit codes, which in turn are divided into 1,239 six-digit codes. FoR has a hierarchical structure and rating might be done at 2-digit level (e.g. 02: Information and Computing Sciences) or 4-digit level (e.g. 0807: Library and Information Studies). There is a volume threshold for rating in code. For instance, in 2018 if the number of outputs of a university in a specific code was below 50, it was not assessed in that code and therefore in the results, it would be marked as 'Not assessed'.

Example is below

080000	INFORMATION AND COMPUTING SCIENCES
080100	Artificial Intelligence and Image Processing
080101	Adaptive Agents and Intelligent Robotics
…	
080700	Library and Information Studies
080701	Aboriginal and Torres Strait Islander Knowledge Management
080702	Health Informatics
080703	Human Information Behaviour
….	

This study analyses the rating of 38 Australian universities in ERA. We calculate a few measures based on the results of ERA 2015 (the last one available) as below:

- Diversification Breadth: the number of FoR codes in which a university has been assessed and received a rating divided by the total number of FoR codes (179, 22 two-digit + 157 four-digit). This shows the ratio of fields in which the university is research active. In other words, it is interpreted as the breadth of research activity in a university.
- Diversification Depth: Total number of FoR codes (2 or 4-digit) codes in which a university has been assessed and received a rating of 4 or 5 divided by the total number of codes in which the university has been assessed. This shows the ratio of fields where a university is very active and has produced high-quality research and/or high research income. In other words, this reflects on the quality of diversification.

- Average of Ranks: Average of all the ratings a university has received.
- Variance of Ranks: Variance of all the ratings a university has received which shows how much disparity the ratings of the university's research in different fields have.

To measure the performance of the university's research, we use some of the InCites' database (Clarivate Analytics) indicators as below. The indicators were obtained for the period 2011-2013 which was the reference period for 2015 ERA. All definitions below, except for h-index, are from Clarivate Analytics)2018(.

- Category Normalized Citation Impact: This is a modification of the category normalized citation impact taking into account the country/region where the institution is based.
- % Documents in Top 1%: Percentage of publications in the top 1% based on citations by category, year, and document type.
- % Documents in Top 10%: Percentage of publications in the top 10% based on citations by category, year, and document type.
- % Highly Cited Papers: Percentage of publications that are assigned as Highly Cited in ESI (top 1% by citations for field and year).
- Citation Impact: The citation impact of a set of documents is calculated by dividing the total number of citations by the total number of publications.
-
- Impact Relative to World: Citation impact of the set of publications as a ratio of the world average.
- % Industry Collaborations: Percentage of publications that have co-authors from industry.
- % International Collaborations: Percentage of publications that have international co-authors.
- H-Index: it is the maximum value of h such that the given author (in this case a university) has published h papers that have each been cited at least h times (Hirsch, 2005).

Moreover, we used QS (Quacquarelli Symonds) World University Ranking of the universities (https://www.topuniversities.com) (as the ranking was available for most of the Australian universities) to find out about their overall performance in the world.

Findings

Table 1 shows the statistics for diversification breadth and depth of Australian universities based on ERA data. Average of ranks of universities in different fields and variance of their ranks are also presented. For instance, University of New South Wales, Sydney (UNSWS) has a very high value of diversification depth (0.92) and a medium level of diversification breadth (0.47), while Charles Sturt University (CSU) has a very low level of diversification depth (0.07) and small amount of diversification breadth (0.26). Average shows a university's overall performance as it is the mean of all its ratings in all fields. The highest average belongs to the Australian National University, ANU (4.49) and the lowest to the University of Notre Dame Australia, UNDA (2.33). The variance shows how the ratings of the university in different fields are distributed. A larger variance value means that the university has a range of low and high ratings. Central Queensland University, CQU has the largest variance (1.79) while the lowest

variance belongs to the University of Queensland, UQ (0.38). This is because UQ was rated 5 in 59 fields, 4 in 52 fields and 3 in 8 fields and received no 2 or 1 ratings. On the other hand, CQU received a rating of 5 in 6 fields, 4 in 5 fields, 3 in 3 fields, 2 in 10 fields, and a 1 in 2 fields.

Table 1. Statistics of diversification breadth and depth of universities in ERA

Uni	*Diversification breadth*	*Diversification depth*	*AVG of ranks*	*Variance of ranks*
ACU	0.15	0.38	3.38	1.08
ANU	0.47	0.92	4.49	0.41
BU	0.11	0.20	2.50	1.35
CQU	0.15	0.42	3.12	1.79
CDU	0.12	0.43	3.29	1.25
CSU	0.26	0.07	2.41	0.72
CUT	0.40	0.45	3.42	0.84
DU	0.40	0.47	3.53	0.75
ECU	0.21	0.26	2.79	1.01
FU	0.42	0.27	3.12	0.72
GU	0.44	0.54	3.63	0.72
JCU	0.35	0.37	3.24	1.07
LTU	0.42	0.49	3.57	1.34
MaU	0.42	0.59	3.79	0.67
MoU	0.63	0.88	4.27	0.46
MuU	0.32	0.28	3.18	0.64
QUT	0.40	0.56	3.71	0.57
RMIT	0.30	0.52	3.63	0.97
SCU	0.16	0.62	3.76	1.15
SUT	0.29	0.40	3.33	1.10
UA	0.50	0.79	4.19	0.67
UC	0.18	0.22	2.97	0.84
UM	0.71	0.89	4.40	0.49
UNE	0.26	0.43	3.48	0.99
UNSWS	0.62	0.77	4.16	0.69
UN	0.43	0.60	3.79	0.92
UNDA	0.07	0.25	2.33	1.56
UQ	0.66	0.93	4.43	0.38
USA	0.30	0.55	3.83	0.97
USQ	0.18	0.28	2.78	1.17
US	0.68	0.84	4.31	0.53
UT	0.40	0.53	3.67	0.83
UTS	0.41	0.63	3.75	0.71
USC	0.13	0.38	3.04	1.29
UWA	0.51	0.67	3.88	0.61
UWS	0.42	0.37	3.32	0.91

UW	0.33	0.68	4.02	0.80
VU	0.18	0.22	2.78	0.98

Table 2 shows citation impact indicators for the universities. UNSWS, for instance, has published 16294 articles in FoR fields between 2011 and 2013 and for that period its category normalized citation impact is 1.45 with 2.12 percent of its articles being among the top 1% and 16.4 percent of its articles being among the top 10%, which are all good values. It has an h-index of 172 and 41.9 percent of its publications included international collaboration.

Table 2. Statistics of citation impact performance of universities

University	*WoS Docs*	*Category Norm. Citation Impact*	*% Docs in Top 1%*	*% Docs in Top 10%*	*% Highly Cited Papers*	*Citation Impact*	*H-Index*	*Impact Relative to World*	*Jrnl Norm. Citation Impact*	*% Industry Collab.*	*% Int. Collab.*
ACU	954	0.91	1.68	14.7	0.63	11.2	45	0.65	1.27	0.10	30.3
ANU	9928	1.51	2.51	18.3	1.83	21.9	150	1.28	1.20	0.87	48.6
BU	711	1.38	1.69	13.6	1.13	18.4	44	1.07	1.19	0.28	43.6
CQU	656	1.12	3.20	18.4	1.98	16.3	48	0.95	1.38	0.46	30.5
CDU	975	1.16	2.26	16.0	1.64	18.8	57	1.10	1.25	0.41	41.2
CSU	1505	0.83	1.20	10.8	0.80	12.3	53	0.72	1.03	0.20	36.4
CUT	5100	1.17	1.84	15.3	1.33	17.5	97	1.03	1.19	1.22	48.0
DU	4570	1.47	2.21	17.8	1.79	22.2	106	1.30	1.26	1.09	41.6
ECU	1142	1.16	1.58	14.3	1.93	19.5	61	1.14	1.13	0.53	47.3
FU	3952	1.24	1.37	12.3	0.94	18.9	82	1.10	1.07	1.32	36.2
GU	5122	1.29	1.80	16.2	1.00	20.1	96	1.18	1.18	0.96	40.3
JCU	3539	1.46	3.14	19.5	2.49	26.0	111	1.52	1.22	1.16	46.9
LTU	3704	1.07	1.48	13.3	1.03	15.2	84	0.89	1.12	0.57	31.8
MaU	4616	1.25	2.12	16.8	1.30	18.1	100	1.06	1.20	0.93	43.2
MoU	18500	1.40	2.13	16.1	1.52	20.5	182	1.20	1.17	1.88	40.3
MuU	1752	1.07	1.83	14.8	1.31	17.2	71	1.01	1.13	2.80	44.5
QUT	4783	1.23	1.55	15.6	1.38	18.8	93	1.10	1.22	1.11	42.4
RMIT	3015	1.18	1.86	16.2	1.13	17.4	77	1.02	1.21	0.73	38.7
SCU	969	1.06	2.27	15.7	1.34	15.6	50	0.91	1.22	0.41	35.9
SUT	2133	1.56	2.44	18.9	1.88	24.5	93	1.43	1.27	1.78	51.5
UA	8798	1.33	2.32	16.3	1.73	20.8	138	1.22	1.19	1.72	42.8
UC	936	1.04	1.50	11.1	0.85	13.0	43	0.76	1.09	0.32	28.3
UM	21337	1.54	2.64	17.8	2.00	23.8	220	1.40	1.17	2.07	43.4
UNE	1455	0.91	1.79	13.0	1.17	13.7	54	0.80	1.09	0.76	34.1
UNSWS	16294	1.45	2.12	16.4	1.56	20.9	172	1.23	1.19	1.72	41.9
UN	5047	1.16	1.60	15.1	1.15	17.8	106	1.04	1.07	1.33	38.6
UNDA	531	1.14	1.13	15.1	0.75	13.5	42	0.79	1.23	0.19	35.8
UQ	19400	1.47	2.54	17.5	1.94	23.5	210	1.37	1.21	1.68	43.9
USA	3849	1.17	1.66	14.9	1.38	18.2	90	1.07	1.15	0.83	39.9
USQ	770	0.91	1.17	10.9	1.17	14.6	50	0.86	1.03	0.91	36.9

US	25941	1.50	2.38	16.3	1.61	20.1	214	1.18	1.26	2.23	43.3
UT	4112	1.31	1.97	15.6	1.48	20.7	98	1.21	1.14	2.07	45.6
UTS	3530	1.28	2.81	18.2	1.78	18.3	99	1.07	1.24	1.22	46.2
USC	558	1.19	1.43	12.0	0.36	14.2	36	0.83	1.15	0.00	36.0
UWA	12973	1.43	2.30	15.7	1.75	21.6	167	1.26	1.18	2.09	47.2
UWS	3321	1.31	2.56	18.0	1.72	18.1	87	1.06	1.26	1.11	43.7
UW	4313	1.28	2.20	16.2	1.46	18.6	103	1.09	1.20	1.23	43.5
VU	1531	1.30	3.46	15.9	3.07	18.6	72	1.09	1.25	0.46	45.1

Correlation matrix of the main variables is presented in Table 3. As expected, Diversification breadth and depth both show statistically significant positive correlations with all the impact variables with different strength levels. The strongest correlation of Diversification breadth is with h-index (0.911) and then with QS ranking (0.799). Its weakest correlation is with the percentage of highly cited papers. While the weakest correlation of Diversification depth is also with the percentage of highly cited papers, its strongest correlations are with QS ranking (0.911) and then h-index (0.793). There was also a positive and relatively strong correlation between Diversification depth and Diversification breadth (0.748). The variance was negatively correlated with almost all of the impact indicators. This might be an indication that universities that have a mix of low and high-quality performance in ERA do not perform as well as those universities that rate consistently high in most of the fields in ERA.

Table 3. Correlation of diversification breadth and depth with citation performance metrics

	Diversification Breadth	*Diversification Depth*	*Average*	*Variance*	*Category Normalized*	*% Documents*	*% Documents*	*% Highly Cited Papers*	*Citation Impact*	*H-Index*	*Impact Relative to*	*% Industry Collaboratio*	*% Int. collaboration*
Diversification Depth	.748												
Average	.790	.977											
Variance	-.823	-.629	-.685										
Category Normalized Citation Impact	.613	.502	.526	-.504									
% Documents in Top 1%	.344	.439	.426	-0.271	.630								
% Documents in Top 10%	.443	.527	.501	-0.307	.704	.864							
% Highly Cited Papers	.336	.357	.339	-0.276	.624	.841	.741						
Citation Impact	.618	.500	.522	-.510	.900	.582	.650	.726					
H-Index	.911	.793	.827	-.761	.754	.531	.618	.561	.785				
Impact Relative to World	.618	.500	.522	-.510	.900	.582	.650	.726	1.000	.785			
% Industry Collaborations	.783	.570	.610	-.680	.591	.394	.423	.439	.651	.805	.651		
% Int. Collaborations	.354	0.219	0.249	-.397	.667	.536	.491	.655	.680	.539	.680	.535	

QS	.799	.911	.905	-.700	.724	.465	.620	.373	.625	.880	.625	.647	.366

Blue: correlation is significant at the 0.01 level; Pink: correlation is significant at 0.05 level.

The next two scatter plots (Figure 1 and 2) show the position of universities in terms of their Diversification depth and Diversification breadth along with some citation impact metrics. The values of diversification breadth and depth are subtracted from their averages in order to draw the diagram. Zero point in both axes is the average. The scatter plots show the universities in four groups including those with high Diversification depth and high diversification breadth (top-right, e.g. UQ: The Uni. of Queensland), high diversification depth and low diversification breadth (top left, e.g. UW: Uni. of Wollongong), low diversification depth and high diversification breadth (bottom right, e.g. FU: Flinders Uni.) and low diversification depth and low diversification breadth (bottom left, e.g. CSU: Charles Sturt Uni.).

In Figure 1, the intensity of the colour indicates category normalized citation impact and the size of the nodes indicate the impact relative to the world. While most of the universities with high relative impact and category normalized citation impact are in the top right section (i.e. high diversification breadth and depth), there are a few exceptions such as James Cook University (JCU) and even Swinburne University of Technology (SUT) while it is located in the bottom left area. Diversification breadth and depth both have moderate positive correlations with these two citation impact indicators. The filled circles represent the Group of Eight, which are Australian leading research-intensive universities including University of Melbourne (UM), the Australian National University (ANU), the University of Sydney (US), the University of Queensland (UQ), the University of Western Australia (UWA), the University of Adelaide (UA), Monash University (MoU) and UNSW Sydney (UNSWS).

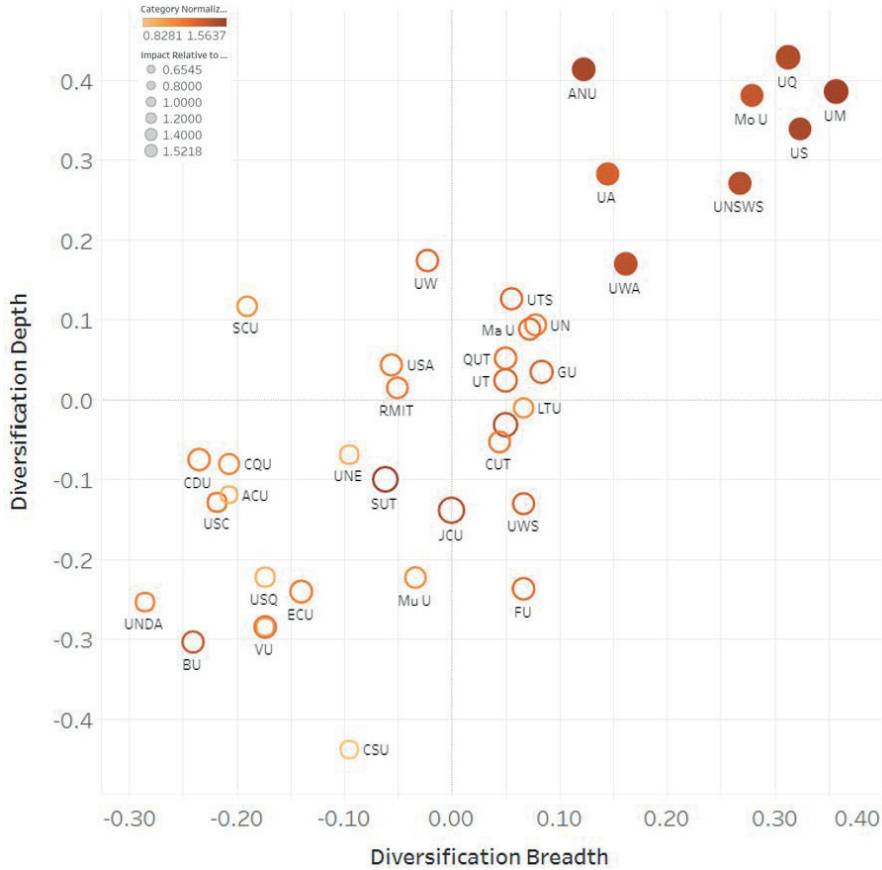


Figure 1. Scatter plot of diversification breadth and depth with category normalized citation impact and impact relative to the world

In Figure 2, the colour intensity represents h-index and the size of nodes indicates the ranking of the university in the QS ranking system. It is very clear that those with high Diversification depth and high Diversification breadth have the largest h-index values. They also clearly perform better in terms of their position in the QS ranking system. H-index has a very strong positive correlation with diversification breadth and a relatively strong correlation with diversification depth. QS Ranking value has a relatively strong correlation with diversification breadth and a very strong correlation with diversification depth.

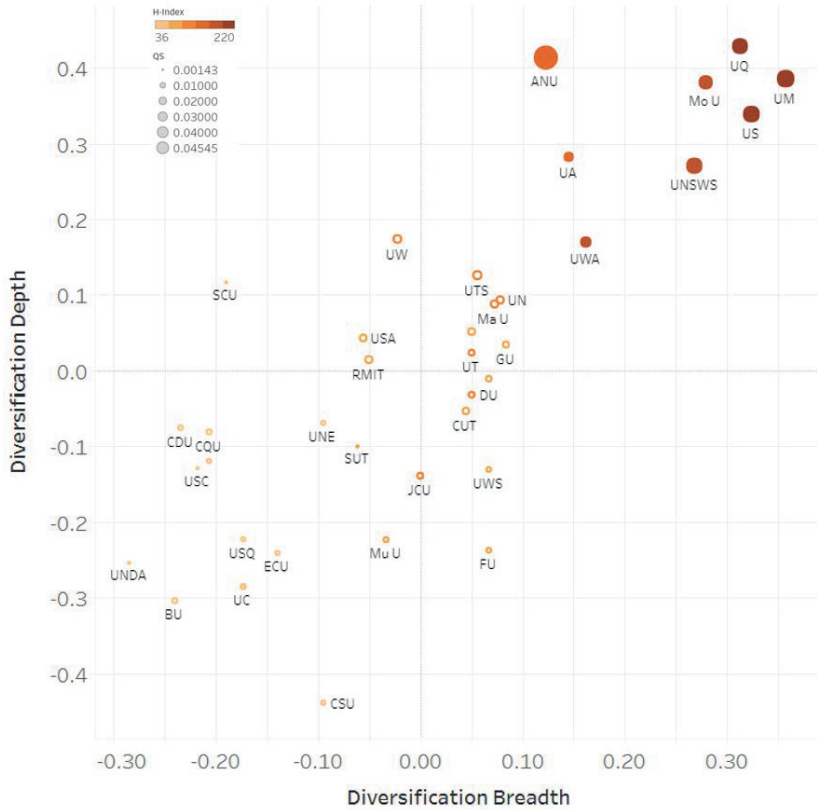


Figure 2. Scatter plot of diversification breadth and depth with h-index and QS ranks

Conclusion

The study used the rating of disciplinary research activities in Australian universities' national evaluation systems called Excellence in Research for Australia (ERA) along with some citation impact indicators form InCites database to look at the relationship between diversification or breadth of research activity in universities and their citation impact performance. The results showed that diversification, measured by counting the number of research fields a university is active in, had moderate to strong correlations with some important citation impact indicators such as university h-index, impact relative to world and category normalized citation impact. It also had a statistically significant correlation with international collaboration. Diversification breadth was also correlated with the ranking of the university in QS ranking of universities.

However, in the case of ERA, diversification breadth was also correlated with diversification depth, i.e. the number of fields in which a university performs above the world average (gaining an ERA rating of 4 or 5). This indicates that universities which are active in a high number of research fields are also more likely to produce high-quality research. Being active in a wide range of research fields, which also mean having more resources either in forms of human resources or equipment, facilitates interdisciplinary research not only internally but also externally, and therefore, leads to higher quality research outputs. University of Melbourne (UM) or Australian National University (ANU) are examples of such cases. Generally, it was

clear in graphs that the Group of Eight had all high diversification breadth (active in many research fields) and high diversification depth (having more high ERA ratings) values (the top right quarter of Figures 1 & 2). On the other hand, regional or teaching-focused universities such as Charles Sturt University (CSU), or Bond University (BU) were mostly in the bottom left quarter of the graph with low diversification breadth and low diversification depth values. These universities in the bottom left quarter did not perform well in terms of citation impact or QS ranking. The other pattern evident from the data was the most universities were either in the high diversification breadth - high diversification depth quarter or in low diversification breadth - low diversification depth quarter. Only a few universities were in the other two quarter where they scored high in one and low in another of the two values of diversification breadth and depth.

The research had a few limitations as it relied on ERA data where an initial threshold applies to the ranking of fields for each university, however, the threshold is not large and if a university has 50 or more papers in a field, it will be assessed in that FoR code.

In a future study, other characteristics of the universities such as their size should be considered to develop normalised measures. Also, a more in-depth analysis of interdisciplinary activities will shed more light on the effects of interdisciplinary research on research performance. This may need to develop new metrics to measure university interdisciplinary activities.

References

Abbott, M., & Doucouliagos, C. (2003). The efficiency of Australian universities: A data envelopment analysis. *Economics of Education Review, 22*(1), 89–97. doi: 10.1016/S0272-7757(01)00068-1

Bonaccorsi, A., & Secondi, L. (2017). The determinants of research performance in European universities: A large scale multilevel analysis. *Scientometrics*, *112*(3), 1147-1178.

Clarivate Analytics (2018). *InCites indicators handbook*. Retrieved April 7, 2019 from http://help.prod-incites.com/inCites2Live/8980-TRS/version/default/part/AttachmentData/data/InCites-Indicators-Handbook%20-%20June%202018.pdf

CL-CSIC (2009). Ranking Web of World Universities, Cybermetrics Lab CSIC. Retrieved April 6, 2017 from http://www.webometrics.info/.

CWTS (2009). Leiden World Ranking, Centre for Science and Technology Studies, University of Leiden. Retrieved April 6, 2017 from http://www.leidenranking.com.

Daraio, C., Bonaccorsi, A., & Simar, L. (2015). Efficiency and economies of scale and specialization in European universities: A directional distance approach. *Journal of Informetrics*, *9*(3), 430-448. doi: 10.1016/j.joi.2015.03.002

Filippini, M., & Lepori, B. (2007). Cost structure, economies of capacity utilization and scope in Swiss higher education institutions. In *Universities and strategic knowledge creation: Specialization and performance in Europe* (pp. 272–304).

Hirsch, J. E. (2005). An index to quantify an individual's scientific research output. *PNAS, 102*(46), 16569-16572. doi: 10.1073/pnas.0507655102

Johnes, G., & Johnes, J. (2009). Higher education institutions' costs and efficiency: Taking the decomposition a further step. *Economics of Education Review, 28*(1), 107–113. doi:10.1016/j.econedurev.2008.02.001

Liu, N. C., & Cheng, Y. (2005). The academic ranking of world universities—methodologies and problems. *Higher Education in Europe, 30*(2), 127–136.

Lopez-Illescas, C., de Moya-Anegon, F., & Moed, H. F. (2011). A ranking of universities should account for differences in their disciplinary specialization. *Scientometrics*, *88*(2), 563–574. doi: 10.1007/s11192-011-0398-6

Moed, H. F., de Moya-Anegón, F., López-Illescas, C., & Visser, M. (2011). Is concentration of university research associated with better research performance? *Journal of Informetrics*, *5*(4), 649-658. doi: 10.1016/j.joi.2011.06.003

Pastor, J. M., & Serrano, L. (2016). The determinants of the research output of universities: specialization, quality and inefficiencies. *Scientometrics*, *109*(2), 1255-1281. doi: 10.1007/s11192-016-2102-3

Pianta, M., & Archibugi, D. (1991). Specialization and size of scientific activities: a bibliometric analysis of advanced countries. *Scientometrics*, *22*(3), 341-358. doi:10.1007/BF02019767

Robinson-Garcia, N., & Jiménez-Contreras, E. (2016). Analyzing the Disciplinary Focus of Universities: Can Rankings Be. *World University Rankings and the Future of Higher Education*, 161-185.

Rousseau, R. (2018). The Australian and New Zealand's Fields of Research (FoR) Codes. ISSI Newsletter. 2018; *14*(3), 59-61.

SCImago (2010). The SIR 2009 World Report. Retrieved April 6, 2017 from http://www.scimagoir.com/.

Thanassoulis, E., Kortelainen, M., Johnes, G., & Johnes, J. (2011). Costs and efficiency of higher education institutions in England: A DEA analysis. *Journal of the Operational Research Society, 62*(7), 1282–1297. doi:10.1057/jors.2010.68

Times QS (2009). Times QS World University Rankings. Retrieved April 6, 2017 http://www.topuniversities.com/.

Wolszczak-Derlacz, J. (2017). An evaluation and explanation of (in) efficiency in higher education institutions in Europe and the US with the application of two-stage semi-parametric DEA. *Research Policy*, *46*(9), 1595-1605. doi: 10.1016/j.respol.2017.07.010

Which are the influential publications in the Web of Science subject categories over a long period of time? CRExplorer software used for big-data analyses in bibliometrics

Andreas Thor[1], Lutz Bornmann[2], Robin Haunschild[3], Loet Leydesdorff[4]

[1] *thor@hft-leipzig.de*
University of Applied Sciences for Telecommunications Leipzig, Gustav-Freytag-Str. 43-45, 04277 Leipzig (Germany).

[2] *bornmann@gv.mpg.de*
Division for Science and Innovation Studies, Administrative Headquarters of the Max Planck Society, Hofgartenstr. 8, 80539 Munich (Germany).

[3] *r.haunschild@fkf.mpg.de*
Max Planck Institute for Solid State Research, Information Service, Heisenbergstrasse 1, 70506 Stuttgart, (Germany).

[4] *loet@leydesdorff.net*
University of Amsterdam, Amsterdam School of Communication Research (ASCoR)P.O. Box 15793, 1001 NG Amsterdam (The Netherlands).

Abstract

What are the landmark papers in scientific disciplines? Which papers are indispensable for scientific progress? These are typical questions which are not only of interest for researchers (who frequently know the answers – or guess to know them), but also for the interested general public. Citation counts can be used to identify very useful papers, since they reflect the wisdom of the crowd; in this case, the scientists using published results for their research. In this study, we identified with recently developed methods for the program CRExplorer landmark publications in nearly all Web of Science subject categories (WoSSCs). These are publications which belong more frequently than other publications during the citing years to the top-‰ in their subject area. The results for three subject categories "Information Science & Library Science", "Computer Science, Information Systems", and "Computer Science, Software Engineering" are exemplarily discussed in more detail. The results for the other WoSSCs can be found online at http://crexplorer.net.

Introduction

Bibliometrics is frequently used in research evaluation. In an overview, Sivertsen (2017) notes that bibliometric indicators are considered in many national research-funding systems in the European Union to measure research performance. Not only researchers themselves, but also science administrators and the public are interested in reports about groundbreaking research from units of assessments (e.g., universities or countries; see e.g., van Noorden, Maher, & Nuzzo, 2014). According to Winnink, Tijssen, and van Raan (2018), the term groundbreaking (or breakthrough) is often used for research (discoveries) with a major impact on future scientific activities. Hollingsworth (2008) considers breakthroughs as very useful to many researchers in targeting future research questions in various scientific fields.

Although breakthroughs are of general interest in science (Orduna-Malea, Martín-Martín, & Delgado López-Cózar, 2018; Schlagberger, Bornmann, & Bauer, 2016), research evaluation focuses – as a rule – on short-time horizons: "the time horizon is 10 years or less, and the focus is on recent past performance, as it is believed to increase the policy relevance, and reduce data collection costs" (Moed, 2017, p. 6). Whereas short-term impact measurements allow statements about the research front, "long-term impact indicates to what extent they eventually succeed in scoring 'triumphs'" (Moed, Burger, Frankfort, & van Raan, 1985, p. 134). The results of Wang (2013) further show that the use of a short citation window (the standard is a minimum of three years) may lead to hasty classifications of papers as high-

impact papers which turn out to be erroneous in the long run (Baumgartner & Leydesdorff, 2014; Leydesdorff, Bornmann, Comins, & Milojević, 2016; Ponomarev, Williams, Hackett, Schnell, & Haak, 2014). The results of Wang, Veugelers, and Stephan (2017) as well as Mairesse and Pezzoni (2018) reveal that novel papers are associated with high citation rates especially in the long run. Garfield, Pudovkin, and Istomin (2002) introduced methods to produce tables with highly-cited papers.

Winnink et al. (2018) studied five algorithms for detecting breakthrough papers. The results point out that the algorithms are powerful tools for tracing breakthrough papers. van Noorden et al. (2014) used traditional citation analyses to identify the most cited publications of all time. They found that about 15,000 papers have more than 1,000 citations and thus seem to be very useful. Marx, Bornmann, Barth, and Leydesdorff (2014) developed the method Reference Publication Year Spectroscopy (RPYS) to detect the origins of research fields or topics. The method is based on counting cited references (instead of citations) to assess the impact of publications on a topic- or field-specific publication set (e.g., climate change, see Marx, Haunschild, Thor, & Bornmann, 2017).[1] The method has already been successfully applied in identifying papers with outstanding performance (Comins & Leydesdorff, 2017, 2018; Thor, Bornmann, Marx, & Mutz, 2018) and landmark patents (Comins, Carmack, & Leydesdorff, 2017).

Thor, Marx, Leydesdorff, and Bornmann (2016) introduced the CRExplorer – a program for undertaking RPYS. In a recent update of the program, Thor et al. (2018) developed an indicator for identifying publications in research fields which are influential over longer periods. In other words, publications (cited references) can be identified which belong to the 10% most-referenced publications in many citing years. In this study, we use a new variant of the indicator to identify publications which belong to the 1‰ (0.1%) most-referenced publications in all citing years between 1980 and 2017 in 205 subject categories (the indicator is named N_TOP0_1+). By focusing on the top-‰, we have identified the exceptionally useful published research in the subject categories between 1980 and 2017. In this paper, the procedure is explained how the publications have been identified. The results for three subject categories are explained in this paper in more detail; the results for all subject categories can be inspected online at http://crexplorer.net.

Methods

Datasets used

We used the Web of Science (WoS, Clarivate Analytics) custom data of the Max Planck Society's in-house database derived from the Science Citation Index Expanded (SCI-E), Social Sciences Citation Index (SSCI), and Arts and Humanities Citation Index (AHCI) produced by Clarivate Analytics (Philadelphia, USA). All records for the papers of the document type "article" published between 1980 and 2017 were exported separately for each WoS subject category (WoSSC). The WoSSCs were ordered by their number of publications from CQ ("Biochemistry & Molecular Biology") with 1,455,479 articles to 9a ("Green & Sustainable Science & Technology") with 3,169 articles (see Leydesdorff, 2006). We required a ratio of linked vs. cited references of at least 0.30 for a WoSSC to be included. The reason is that only WoSSCs with sufficient references covered by the WoS should be considered in the analyses. In total, 205 WoSSCs were considered.

[1] Within a set of publications, the number of references ("cited references") and the number of citations (obtained by the publications) is identical – provided that all cited references can be matched with the publications in the set.

Indicator used

We are interested in those cited references which have been cited disproportionally more frequently in the citing years than other cited references in the dataset. To this end, for each cited reference we count the number of citing years where the cited reference was cited extraordinarily frequently.

For each citing year, all *n* cited references have been sorted in descending order based on their citation counts in the citing year. We then identified the citation count *c* of the cited reference at rank (1+*n*/1000), i.e., the cited reference that follows the first (top) 0.1% cited references. For example, for *n*=10.000 cited references we determined the number of citations of the cited reference at rank #11. All cited references with a citation count greater than *c* are then considered as "top-cited reference" in the citing year if their citation count is additionally above the average of the expected citation count (see Thor et al., 2018, for details on the sequence computation). The metric N_TOP0_1+ is the number of citing years where the cited reference is a "top-cited reference".

The total number of papers and the citation rates in the WoSSCs considered here are very different. Consequently, the number of citing years for belonging to the top cited publications is different in the sets (as the results online at http://crexplorer.net demonstrate). This should be considered in the interpretation of the results. It should be taken into consideration additionally that the citing papers are frequently assigned to more than a single WoSSC. Thus, the same highly cited publications (presented online) can occur in different WoSSCs.

CRExplorer script

The following CRExplorer script was used to perform the RPYS and filter for exceptionally highly referenced publications for each of the WoSSCs:

```
set(n_pct_range: 2, median_range: 2)
importFile(file: "xx_wos.txt", type: "WOS",
RPY: [1900, 2015, false], PY: [1980, 2017, false], maxCR: 0)
info()
cluster(threshold: 0.75, volume: true, page: true, DOI: false)
merge()
exportFile(file: "xx_wos.rpys_CR.csv", type: "CSV_CR",
     sort: ["N_TOP0_1_Plus DESC", "N_CR DESC"],
filter: { it.N_TOP0_1_Plus >= 10 } )
```

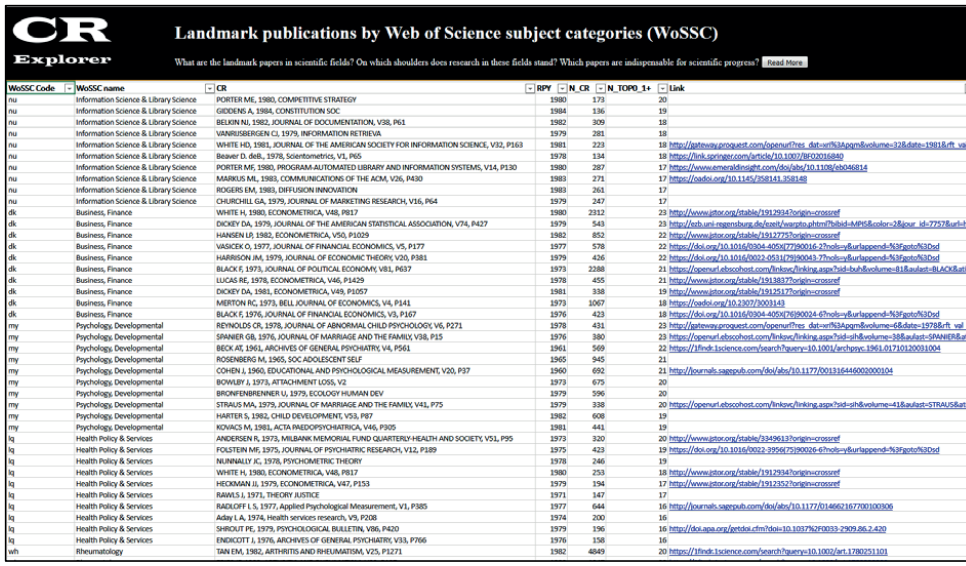
Listing 1: CRExplorer script to perform RPYS and filter for cited references with an indicator value of at least 10 for N_TOP0_1+

Two neighboring years are included in the calculation of the advanced indicators via the set options. Thus, not only the focal years are considered in the calculation, but also neighboring years to increase the case numbers for the analyses. The file name "xx_wos.txt" has to be adjusted for each WoSSC in the importFile function. The PY option ensures that only papers published between 1980 and 2017 are included. The RPY option guarantees that only cited references published between 1900 and 2015 are included. We expect no exceptionally highly referenced papers before 1900. We also expect that cited references published after 2015 did not have enough time to become exceptionally highly referenced during many citing years. The clustering and merging of variants of the same cited reference in the dataset is done with the Levenshtein threshold of 0.75 including volume and page; but not the DOI in the cited references' information (Thor et al., 2016). The file name "xx_wos.rpys_CR.csv" in the exportFile function has to be adjusted for each WoSSC. In addition, this function filters for

cited references with an indicator value of at least 10 (number of citing years where the cited reference is a "top-cited reference") and sorts the results according to the indicator value and the number of cited references before writing the cited references into the csv file. The value of 10 is adjusted to a lower one if cited references in some WoSSCs do not achieve large enough indicator values. For the WoSSCs with many papers and many cited references variants, we needed 382 GB of main memory (RAM).

Results

The identified landmark papers for nearly all WoSSC can be inspected online at http://crexplorer.net (see Figure 1).



CR Explorer

Landmark publications by Web of Science subject categories (WoSSC)

What are the landmark papers in scientific fields? On which shoulders does research in these fields stand? Which papers are indispensable for scientific progress? Read More

WoSSC Code	WoSSC name	CR	RPY	N_CR	N_TOPO_1+	Link
nu	Information Science & Library Science	PORTER ME, 1980, COMPETITIVE STRATEGY	1980	173	20	
nu	Information Science & Library Science	GIDDENS A, 1984, CONSTITUTION SOC	1984	136	19	
nu	Information Science & Library Science	BELKIN NJ, 1982, JOURNAL OF DOCUMENTATION, V38, P61	1982	309	18	
nu	Information Science & Library Science	VANRIJSBERGEN CJ, 1979, INFORMATION RETRIEVA	1979	281	18	
nu	Information Science & Library Science	WHITE HD, 1981, JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE, V32, P163	1981	223	18	http://gateway.proquest.com/openurl?res_dat=xri%3Apqrm&volume=32&date=1981&rft_val_f
nu	Information Science & Library Science	Beaver D deB., 1978, Scientometrics, V1, P65	1978	134	18	https://link.springer.com/article/10.1007/BF02016840
nu	Information Science & Library Science	PORTER MF, 1980, PROGRAM-AUTOMATED LIBRARY AND INFORMATION SYSTEMS, V14, P130	1980	287	17	https://www.emeraldinsight.com/doi/abs/10.1108/eb046814
nu	Information Science & Library Science	MARKUS ML, 1983, COMMUNICATIONS OF THE ACM, V26, P430	1983	271	17	https://oadoi.org/10.1145/358141.358148
nu	Information Science & Library Science	ROGERS EM, 1983, DIFFUSION INNOVATION	1983	261	17	
nu	Information Science & Library Science	CHURCHILL GA, 1979, JOURNAL OF MARKETING RESEARCH, V16, P64	1979	247	17	
dk	Business, Finance	WHITE H, 1980, ECONOMETRICA, V48, P817	1980	2312	23	http://www.jstor.org/stable/1912934?origin=crossref
dk	Business, Finance	DICKEY DA, 1979, JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION, V74, P427	1979	543	23	http://gzb.uni-regensburg.de/ezeit/warpto.phtml?bibid=MPIS&color=2&jour_id=7757&url=htt
dk	Business, Finance	HANSEN LP, 1982, ECONOMETRICA, V50, P1029	1982	852	22	http://www.jstor.org/stable/1912757?origin=crossref
dk	Business, Finance	VASICEK O, 1977, JOURNAL OF FINANCIAL ECONOMICS, V5, P177	1977	578	22	https://doi.org/10.1016/0304-405X(77)90016-2?mobs=y&urlappend=%3Fgoto%3Dsd
dk	Business, Finance	HARRISON JM, 1979, JOURNAL OF ECONOMIC THEORY, V20, P381	1979	426	22	https://doi.org/10.1016/0022-0531(79)90043-7?mobs=y&urlappend=%3Fgoto%3Dsd
dk	Business, Finance	BLACK F, 1973, JOURNAL OF POLITICAL ECONOMY, V81, P637	1973	2288	21	https://openurl.ebscohost.com/linksvc/linking.aspx?sid=bufk&volume=81&aulast=BLACK&atitle
dk	Business, Finance	LUCAS RE, 1978, ECONOMETRICA, V46, P1429	1978	455	21	http://www.jstor.org/stable/1913837?origin=crossref
dk	Business, Finance	DICKEY DA, 1981, ECONOMETRICA, V49, P1057	1981	338	19	http://www.jstor.org/stable/1912517?origin=crossref
dk	Business, Finance	MERTON RC, 1973, BELL JOURNAL OF ECONOMICS, V4, P141	1973	1067	18	https://oadoi.org/10.2307/3003143
dk	Business, Finance	BLACK F, 1976, JOURNAL OF FINANCIAL ECONOMICS, V3, P167	1976	423	18	https://doi.org/10.1016/0304-405X(76)90024-6?mobs=y&urlappend=%3Fgoto%3Dsd
my	Psychology, Developmental	REYNOLDS CR, 1978, JOURNAL OF ABNORMAL CHILD PSYCHOLOGY, V6, P271	1978	431	23	http://gateway.proquest.com/openurl?res_dat=xri%3Apqrm&volume=6&date=1978&rft_val_fm
my	Psychology, Developmental	SPANIER GB, 1976, JOURNAL OF MARRIAGE AND THE FAMILY, V38, P15	1976	380	23	https://openurl.ebscohost.com/linksvc/linking.aspx?sid=sih&volume=38&aulast=SPANIER&atitl
my	Psychology, Developmental	BECK AT, 1961, ARCHIVES OF GENERAL PSYCHIATRY, V4, P561	1961	569	22	https://1find.1science.com/search?query=10.1001/archpsyc.1961.01710120031004
my	Psychology, Developmental	ROSENBERG M, 1965, SOC ADOLESCENT SELF	1965	945	21	
my	Psychology, Developmental	COHEN J, 1960, EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT, V20, P37	1960	692	21	http://journals.sagepub.com/doi/abs/10.1177/001316446002000104
my	Psychology, Developmental	BOWLBY J, 1973, ATTACHMENT LOSS, V2	1973	675	20	
my	Psychology, Developmental	BRONFENBRENNER U, 1979, ECOLOGY HUMAN DEV	1979	596	20	
my	Psychology, Developmental	STRAUS MA, 1979, JOURNAL OF MARRIAGE AND THE FAMILY, V41, P75	1979	338	20	https://openurl.ebscohost.com/linksvc/linking.aspx?sid=sih&volume=41&aulast=STRAUS&atitle
my	Psychology, Developmental	HARTER S, 1982, CHILD DEVELOPMENT, V53, P87	1982	608	19	
my	Psychology, Developmental	KOVACS M, 1981, ACTA PAEDOPSYCHIATRICA, V46, P305	1981	441	19	
lq	Health Policy & Services	ANDERSEN R, 1973, MILBANK MEMORIAL FUND QUARTERLY-HEALTH AND SOCIETY, V51, P95	1973	320	20	http://www.jstor.org/stable/3349613?origin=crossref
lq	Health Policy & Services	FOLSTEIN MF, 1975, JOURNAL OF PSYCHIATRIC RESEARCH, V12, P189	1975	423	19	https://doi.org/10.1016/0022-3956(75)90026-6?mobs=y&urlappend=%3Fgoto%3Dsd
lq	Health Policy & Services	NUNNALLY JC, 1978, PSYCHOMETRIC THEORY	1978	246	19	
lq	Health Policy & Services	WHITE H, 1980, ECONOMETRICA, V48, P817	1980	253	18	http://www.jstor.org/stable/1912934?origin=crossref
lq	Health Policy & Services	HECKMAN JJ, 1979, ECONOMETRICA, V47, P153	1979	194	17	http://www.jstor.org/stable/1912352?origin=crossref
lq	Health Policy & Services	RAWLS J, 1971, THEORY JUSTICE	1971	147	17	
lq	Health Policy & Services	RADLOFF L S, 1977, Applied Psychological Measurement, V1, P385	1977	644	16	http://journals.sagepub.com/doi/abs/10.1177/014662167700100306
lq	Health Policy & Services	Aday L A, 1974, Health services research, V9, P208	1974	200	16	
lq	Health Policy & Services	SHROUT PE, 1979, PSYCHOLOGICAL BULLETIN, V86, P420	1979	196	16	http://doi.apa.org/getdoi.cfm?doi=10.1037%2F0033-2909.86.2.420
lq	Health Policy & Services	ENDICOTT J, 1976, ARCHIVES OF GENERAL PSYCHIATRY, V33, P766	1976	158	16	
wh	Rheumatology	TAN EM, 1982, ARTHRITIS AND RHEUMATISM, V25, P1271	1982	4849	20	https://1find.1science.com/search?query=10.1002/art.1780251101

Figure 1. Online presentation of the landmark papers

In the following we focus exemplarily on three WoSSCs and explain the results in more detail. We selected WoSSCs which we are able to interpret based on our own field-specific expertises. Table 1 shows the results for the WoSSC "Information Science & Library Science". Five cited publications are listed exemplarily with the most citing years in which the publication belongs to the top-‰. Two publications in the table are basic works on information retrieval (Belkin, Oddy, & Brooks, 1982; Van Rijsbergen, 1979). Three of the five publications in the table are not primarily contributions to the library and information science (LIS) field: Michael Porter's (1980) book is one of his contributions to the field of business economics. In later work, Porter (1990) became specifically known for cluster analysis in the follow-up book entitled "The Competitive Advantage of Nations." Anthony Giddens' (1984) book entitled "The Constitution of Society" is the *locus classicus* of Giddens' "structuration theory" in sociology. Both this book and Porter (1980) are well known and intensively used in communications among non-specialists. Both books are theoretical, but oriented towards application (without providing a methodology). White and Griffith (1981) introduced author-co-citation analysis (ACA) in LIS and Science &

Technology Studies. ACA became thereafter a widely used technique. It is primarily a statistical method, but it can also be used in qualitative analysis.

Table 1. Most exceptionally referenced cited references in the WoSSC "Information Science & Library Science".

RPY	CR	N_CR	N_TOP0_1+
1980	Porter, M. E.: *Competitive Strategy: Techniques for Analyzing Industries and Competitors*. Free Press	173	20
1984	Giddens, A.: *The Constitution* of *Society*. Outline of the Theory of Structuration. Polity Press	136	19
1982	*Belkin*, *N. J.*, Oddy, R. N. and Brooks, H.: ASK for Information Retrieval: Part I. Background and Theory. *Journal of Documentation*, *38*(2), *61*-71	309	18
1979	Van Rijsbergen, C.J.: *Information Retrieval*. Unpublished PhD thesis, Department of Computing Science, University of Glasgow	281	18
1981	White, H. D., & Griffith, B. C.: Author Cocitation - a Literature Measure of Intellectual Structure. *Journal of the American Society for Information Science, 32*(3), 163-171	223	18

Notes. RPY=Reference publication year; CR=Cited reference; N_CR=Number of cited references; N_TOP0_1+=Number of citing years in which the publication belongs to the top-‰.

Table 2 shows the results for the WoSSC "Computer Science, Information Systems". The three papers "A Method for obtaining digital Signatures and public-key Cryptosystems" (Rivest, Shamir, & Adleman, 1978), "A public-key Cryptosystem and a Signature Scheme based on discrete Logarithms" (ElGamal, 1985), and "New Directions in Cryptography" (Diffie & Hellman, 2006) describe fundamental algorithms for data encryption and digital signatures. These algorithms are important for secure (i.e., encrypted) data transmission over the Internet. The idea of an asymmetric cryptosystem based on public and private keys (that can be exchanged securely) is used in current software such as PGP. Rivest et al. (1978) also received the ACM Turing award (the "Nobel prize for computer science") for their work. The book by Garey and Johnson (1979) "Computers and Intractability: a Guide to the Theory of NP-Completeness" gives an introduction to computational complexity, a fundamental concept in theoretical computer science. The book is well-known for its extensive list of NP-complete problems, i.e., problems where an efficient solution (i.e., in polynomial time) does not yet exist. Especially in the era of big data, efficient software algorithms (besides large clusters of hardware components) are a cornerstone of many web applications. "The Theory of error-correcting Codes" (MacWilliams & Sloane, 1977) is an influencing book on information theory and coding theory. It describes approaches for the reliable transmission of data over unreliable communication channels, e.g., when multiple mobile phones interfere with each other on the same WiFi network.

Table 2. Most exceptionally referenced cited references in the WoSSC "Computer Science, Information Systems".

RPY	CR	N_CR	N_TOP0_1+
1978	Rivest, R. L., Shamir, A., & Adleman, L. (1978). A Method for obtaining digital Signatures and public-key Cryptosystems. *J Commun. ACM, 21*(2), 120-126	862	21
1979	Garey, M. R. & Johnson, D. S. (1979). *Computers and Intractability: A Guide to the Theory of NP-completeness.* W. H. Freeman	1137	19
1977	MacWilliams, F. J., & Sloane, N. J. A. (1977). *The Theory of error Correcting Codes.* North-Holland Publishing Company	689	19
1985	ElGamal, T. (1985). *A public key Cryptosystem and a Signature Scheme Based on discrete Logarithms.* Paper presented at the Workshop on the Theory and Application of Cryptographic Techniques, Berlin, Heidelberg	503	19
1976	Diffie, W., & Hellman, M. (2006). New Directions in Cryptography *J IEEE Trans. Inf. Theor, 22*(6), 644-654	878	18

Notes. RPY=Reference publication year; CR=Cited reference; N_CR=Number of cited references; N_TOP0_1+=Number of citing years in which the publication belongs to the top-‰.

The results for the WoSSC "Computer Science, Software Engineering" are reported in Table 3. The first two cited references are the in area of theoretical computer science. The book by Garey and Johnson (1979) has already been described since it also appears in the top list of "Computer Science, Information Systems". The paper "Maintaining Knowledge about temporal Intervals" (Allen, 1983) introduces a calculus for temporal reasoning. This is important for software or robots using artificial intelligence where the concept of time (i.e., when things happen) is important. The two papers "Recursively generated B-spline Surfaces on arbitrary topological Meshes" (Catmull & Clark, 1978) and "Theory of Edge Detection" (Marr, Hildreth, & Brenner, 1980) are in the area of computer graphics. The technique of B-spline surfaces is used in computer graphics to create smooth surfaces. This is, for example, important in 3D video games to generate realistically looking objects. Edge detection is a core task in processing digital images to detect and extract features (e.g., objects) in digital images. This is particularly important in computed tomography technique (CT) to detect objects of interest, e.g., arteries. Weiser (1984) introduced the concept of "Program slicing", a method for automatically decomposing programs into so-called slices. The decomposition can be used for efficient finding of errors (debugging) but also for software maintenance and optimization. Though the concept has been significantly extended over the years, it is still a fundamental concept in professional software engineering.

Table 3. Most exceptionally referenced cited references in the WoSSC "Computer Science, Software Engineering".

RPY	CR	N_CR	N_TOP0_1+
1979	Garey, M. R. & Johnson, D. S. (1979). *Computers and Intractability: A Guide to the Theory of NP-completeness.* W. H. Freeman	867	19
1983	Allen, J. F. (1983). Maintaining knowledge about temporal Intervals. *J Commun. ACM, 26*(11), 832-843	231	19
1978	Catmull, E., & Clark, J. (1978). Recursively generated B-spline Surfaces on arbitrary topological Meshes. *Computer-Aided Design, 10*(6), 350-355	364	18
1980	Marr, D., Hildreth, E., & Brenner, S. (1980). Theory of Edge Detection. *207*(1167), 187-217	206	18
1984	Weiser, M. (1984). Program slicing. *IEEE Transactions on Software Engineering, SE-10*(4), 439-449	351	17

Notes. RPY=Reference publication year; CR=Cited reference; N_CR=Number of cited references; N_TOP0_1+=Number of citing years in which the publication belongs to the top-‰.

Discussion

What are the landmark papers in scientific fields? Which papers would be indispensable for scientific progress? These are typical questions which are not only of interest for researchers (who frequently know the answers – or are supposed to know them), but also for the general public (e.g., science journalists). Citation counts are often used to identify very useful papers, since they reflect the wisdom of the crowd; in this case, the many scientists citing the published results in their own papers. The problem with today's research evaluation processes is, however, that they focus on rather recent years (the last few years) to assess the recent developments. This focus might be able to identify research at the research front which is short-term oriented, but neglect research which appears successful in the long run. Extreme representatives of delayed recognition are so-called "sleeping beauties" which are not or scarcely cited during many years, but are heavily cited after a decade or so. These papers become useful only many years after the research has been finished.

In this study, we identified landmark publications in 205 WoSSCs with recently developed methods for the program CRExplorer. These are publications which belong more frequently than other publications to the top-‰ in their subject category across the citing years. In this paper, the results for the three WoSSCs "Information Science & Library Science", "Computer Science, Information Systems", and "Computer Science, Software Engineering" have been discussed in more detail. The results for nearly all WoSSCs can be found online (see http://crexplorer.net). It was only possible with a very powerful computer to generate the results for very large WoSSCs in our dataset. Since most users of the CRExplorer do not have these computers for undertaking cited references analyses, we deem it useful for researchers in various fields, science administrators, science journalists, and other people from the general public to have access to these landmark papers' lists.

The identification of very useful research based on citations (or cited references) is based on the premise that citations measure usefulness. Recent research suggests that citations reflect

"usefulness" which supports the use of citations in science studies and evaluation practices (Wang, 2014). However, citations are not able to reflect all influences which were useful for extraordinary research (the later landmark papers). It is especially relevant for extraordinary research to be influenced by many channels to receive this specific status (MacRoberts & MacRoberts, 2017). Another problem is the incompleteness of many reference lists: "No one who has read J. D. Watson's (1968) personal account of the discovery of the structure of DNA can ever accept that the six references listed at the end of the famous Watson and Crick 1953 paper in *Nature* reflect the influence on their discovery … It is also clear from all accounts that, by 1952, it was the informal level of communication that was important. It was what the scientists were doing on the moving edge of research/speculation that was important to Watson and Crick, and they made every effort to get that information. Clearly, the Watson and Crick paper, similar to all scientific papers, is a 'misrepresentation' of what scientists actually do" (MacRoberts & MacRoberts, 2017, p. 475).

Our generated lists should only be used as hints to possible landmark publications. Users of the lists should be experts in the fields (or should consult experts) who can compare the results with their own reception of landmark papers. For example, in the "Information Science & Library Science" field, the results seem counter-intuitive (against the backdrop of our expert knowledge). One would not expect Porter (1980) and Giddens (1984) to head the ranks. However, one should consider in the interpretation of the results presented in this paper and online at http://crexplorer.net that only up to ten classic papers are presented and many others follow which are (somewhat) lower ranked. The user of our lists should be aware of the fact that the quality of specific publications is not necessarily reflected in high citation counts (Marx & Bornmann, 2010).

Acknowledgments

The bibliometric data used in this paper are from the Max Planck Society's in-house database. The database is developed and maintained in cooperation with the Max Planck Digital Library (MPDL, Munich). It is derived from the Science Citation Index Expanded (SCI-E), Social Sciences Citation Index (SSCI), Arts and Humanities Citation Index (AHCI) prepared by Clarivate Analytics, formerly the IP & Science business of Thomson Reuters (Philadelphia, Pennsylvania, USA).

References

Allen, J. F. (1983). Maintaining knowledge about temporal intervals. *J Commun. ACM, 26*(11), 832-843. doi: 10.1145/182.358434.

Baumgartner, S. E., & Leydesdorff, L. (2014). Group-based trajectory modeling (GBTM) of citations in scholarly literature: Dynamic qualities of "transient" and "sticky knowledge claims". *Journal of the Association for Information Science and Technology, 65*(4), 797-811. doi: 10.1002/asi.23009.

Belkin, N. J., Oddy, R. N., & Brooks, H. M. (1982). Ask for Information-Retrieval .1. Background and Theory. *Journal of Documentation, 38*(2), 61-71. doi: DOI 10.1108/eb026722.

Catmull, E., & Clark, J. (1978). Recursively generated B-spline surfaces on arbitrary topological meshes. *Computer-Aided Design, 10*(6), 350-355. doi: 10.1016/0010-4485(78)90110-0.

Comins, J. A., Carmack, S. A., & Leydesdorff, L. (2017). Patent Citation Spectroscopy (PCS): Algorithmic retrieval of landmark patents. *Journal of Informetrics, 12*(4), 1223-1231.

Comins, J. A., & Leydesdorff, L. (2017). Citation algorithms for identifying research milestones driving biomedical innovation. *Scientometrics, 110*(3), 1495–1504. doi: 10.1007/s11192-016-2238-1.

Comins, J. A., & Leydesdorff, L. (2018). Data-mining the Foundational Patents of Photovoltaic Materials: An application of Patent Citation Spectroscopy. *Journal of Scientometric Research, 7*(2), 79-83.

Diffie, W., & Hellman, M. (2006). New directions in cryptography. *J IEEE Trans. Inf. Theor, 22*(6), 644-654. doi: 10.1109/tit.1976.1055638.

ElGamal, T. (1985). *A Public Key Cryptosystem and a Signature Scheme Based on Discrete Logarithms.* Paper presented at the Workshop on the Theory and Application of Cryptographic Techniques, Berlin, Heidelberg.

Garey, M. R., & Johnson, D. S. (1979). *Computers and Intractability: A Guide to the Theory of NP-completeness*: W. H. Freeman.

Garfield, E., Pudovkin, A. I., & Istomin, V. S. (2002). Algorithmic citation-linked historiography - Mapping the literature of science. *Asist 2002: Proceedings of the 65th Asist Annual Meeting, Vol 39, 2002, 39*, 14-24. doi: DOI 10.1002/meet.1450390102.

Giddens, A. (1984). *The Constitution of Society: Outline of the Theory of Structuration*. Cambridge: Polity Press.

Hollingsworth, J. R. (2008). Scientific discoveries: an institutionalist and path-dependent perspective. In C. Hannaway (Ed.), *Biomedicine in the Twentieth Century: Practices, Policies, and Politics, volume 72 of Biomedical and Health Research* (pp. 317–353). Bethesda, MD, USA: National Institutes of Health.

Leydesdorff, L., Bornmann, L., Comins, J. A., & Milojević, S. (2016). Citations: Indicators of Quality? The Impact Fallacy. *Frontiers in Research Metrics and Analytics, 1*(1). doi: 10.3389/frma.2016.00001.

Leydesdorff, T. (2006). Can scientific journals be classified in terms of aggregated journal-journal citation relations using the Journal Citation Reports? *Journal of the American Society for Information Science and Technology, 57*(5), 601-613.

MacRoberts, M. H., & MacRoberts, B. R. (2017). The mismeasure of science: Citation analysis. *Journal of the Association for Information Science and Technology, 69*(3), 474-482. doi: 10.1002/asi.23970.

MacWilliams, F. J., & Sloane, N. J. A. (1977). *The Theory of Error Correcting Codes*: North-Holland Publishing Company.

Mairesse, J., & Pezzoni, M. (2018). Novelty in Science: The impact of French physicists' novel articles. In P. Wouters (Ed.), *Proceedings of the science and technology indicators conference 2018 Leiden "Science, Technology and Innovation indicators in transition"*. Leider, the Netherlands: University of Leiden.

Marr, D., Hildreth, E., & Brenner, S. (1980). Theory of edge detection. *207*(1167), 187-217. doi: doi:10.1098/rspb.1980.0020.

Marx, W., & Bornmann, L. (2010). How accurately does Thomas Kuhn's model of paradigm change describe the transition from a static to a dynamic universe in cosmology? A historical reconstruction and citation analysis. *Scientometrics, 84*(2), 441-464.

Marx, W., Bornmann, L., Barth, A., & Leydesdorff, L. (2014). Detecting the historical roots of research fields by reference publication year spectroscopy (RPYS). *Journal of the Association for Information Science and Technology, 65*(4), 751-764. doi: 10.1002/asi.23089.

Marx, W., Haunschild, R., Thor, A., & Bornmann, L. (2017). Which early works are cited most frequently in climate change research literature? A bibliometric approach based on Reference Publication Year Spectroscopy. *Scientometrics, 110*(1), 335-353. doi: 10.1007/s11192-016-2177-x.

Moed, H. F. (2017). *Applied Evaluative Informetrics*. Heidelberg, Germany: Springer.

Moed, H. F., Burger, W. J. M., Frankfort, J. G., & van Raan, A. F. J. (1985). The use of bibliometric data for the measurement of university research performance. *Research Policy, 14*(3), 131-149.

Orduna-Malea, E., Martín-Martín, A., & Delgado López-Cózar, E. (2018). Classic papers: using Google Scholar to detect the highly-cited documents. In P. Wouters (Ed.), *Proceedings of the science and technology indicators conference 2018 Leiden "Science, Technology and Innovation indicators in transition"*. Leider, the Netherlands: University of Leiden.

Ponomarev, I. V., Williams, D. E., Hackett, C. J., Schnell, J. D., & Haak, L. L. (2014). Predicting highly cited papers: A Method for Early Detection of Candidate Breakthroughs. *Technological Forecasting and Social Change, 81*(0), 49-55. doi: 10.1016/j.techfore.2012.09.017.

Porter, M. E. (1980). *Competitive Strategy: Techniques for Analyzing Industries and Competitors*. New York: Free Press.
Porter, M. E. (1990). *Competitive Advantage of Nations: Creating and Sustaining Superior Performance*. New York: Free Press.
Rivest, R. L., Shamir, A., & Adleman, L. (1978). A method for obtaining digital signatures and public-key cryptosystems. *J Commun. ACM, 21*(2), 120-126. doi: 10.1145/359340.359342.
Schlagberger, E. M., Bornmann, L., & Bauer, J. (2016). At what institutions did Nobel laureates do their prize-winning work? An analysis of biographical information on Nobel laureates from 1994 to 2014. *Scientometrics, 109*(2), 723–767.
Sivertsen, G. (2017). Problems and considerations in the design of bibliometric indicators for national performance based research funding systems *Proceedings of the Science, Technology, & Innovation Indicators Conference "Open indicators: innovation, participation and actor-based STI indicators*. Paris, France.
Thor, A., Bornmann, L., Marx, W., & Mutz, R. (2018). Identifying single influential publications in a research field: New analysis opportunities of the CRExplorer. *Scientometrics, 116*(1), 591–608.
Thor, A., Marx, W., Leydesdorff, L., & Bornmann, L. (2016). Introducing CitedReferencesExplorer (CRExplorer): A program for reference publication year spectroscopy with cited references standardization. *Journal of Informetrics, 10*(2), 503-515. doi: 10.1016/j.joi.2016.02.005.
van Noorden, R., Maher, B., & Nuzzo, R. (2014). The Top 100 Papers. *Nature, 514*(7524), 550-553.
Van Rijsbergen, C. J. (1979). *Information Retrieval. Unpublished PhD thesis*. Glasgow, UK: Department of Computing Science, University of Glasgow.
Wang, J. (2013). Citation time window choice for research impact evaluation. *Scientometrics, 94*(3), 851-872. doi: 10.1007/s11192-012-0775-9.
Wang, J. (2014). Unpacking the Matthew effect in citations. *Journal of Informetrics, 8*(2), 329-339. doi: 10.1016/j.joi.2014.01.006.
Wang, J., Veugelers, R., & Stephan, P. E. (2017). Bias Against Novelty in Science: A Cautionary Tale for Users of Bibliometric Indicators. *Research Policy, 46*(8), 1416-1436.
Weiser, M. (1984). Program slicing. *IEEE Transactions on Software Engineering, SE-10*(4), 439-449.
White, H. D., & Griffith, B. C. (1981). Author Cocitation - a Literature Measure of Intellectual Structure. *Journal of the American Society for Information Science, 32*(3), 163-171. doi: DOI 10.1002/asi.4630320302.
Winnink, J. J., Tijssen, R. J. W., & van Raan, A. F. J. (2018). Searching for new breakthroughs in science: How effective are computerised detection algorithms? *Technological Forecasting and Social Change*. doi: 10.1016/j.techfore.2018.05.018.

The influence of corresponding authorship on the impact of collaborative publications: a study on Brazilian institutions (2003-2015)

Maria Cláudia Cabrini Grácio[1], Ely Francina Tannuri de Oliveira[2],

Zaida Chinchilla-Rodríguez[3] and Henk F. Moed[4]

[1]*cabrini.gracio@unesp.br*
São Paulo State University, Av. Hygino Muzzi Filho 737, 17525-900, Marília (Brazil)

[2]*etannuri@gmail.com*
São Paulo State University, Av. Hygino Muzzi Filho 737, 17525-900, Marília (Brazil)

[3]*zaida.chinchilla@csic.es*
Spanish National Research Council, Calle de Alfonso XII, 54, 28014, Madrid (Spain)

[4]*hf.moed@gmail.com*
Sapienza University of Rome, Piazzale Aldo Moro, 5, 00185, Rome (Italy)

Abstract

This paper analyses the influence of a Brazilian institution delivering the corresponding author on its scientific citation impact, distinguishing between its collaborative papers with foreign institutions and those resulting from national collaboration. We retrieved from Scopus database a total of 607,454 Brazilian documents for all 443 Brazilian institutions with at least 100 documents published from 2003 to 2015. We evaluated the difference between the normalized citation impact as corresponding author and that as non-corresponding author, applying paired t-tests both for international and for national collaboration. As result, for international collaboration, it was observed that the normalized citation impact achieved by Brazilian institutions depends upon corresponding author status, and that, in case of non-corresponding authorship, the impact shows a significant benefit when the paper has a corresponding author from a foreign institution. In national collaboration, the institutions benefit as non-corresponding author, although the difference is small to influence the practice of the institutions' scientific policies. Thus, the indicator of corresponding author provides additional information relevant to Brazilian institutions in international collaboration, but not in national institutional co-authorship, which is more influenced by the institution's recognized scientific tradition and publishing strategies/practices.

Introduction

Several studies show that scientific collaboration contributes to the broader recognition and impact of the science produced, especially those involving international partnerships (Glänzel & Lange, 2002; Persson, Glänzel & Dannell, 2004; Alonso Arroyo et al., 2016). At an extramural level, mainly between countries, scientific collaboration has become an essential practice to reach the critical mass that consolidates and facilitates the internationalization of new knowledge and the analysis of the science produced (Katz & Martin, 1997; Glänzel, 2003).

An indicator of the scientific collaboration activity is co-authorship. Among its advantages, it is noted that it is based on objective data and is validated in earlier studies; it represents an accessible and user-friendly methodology for quantifying collaboration; it enables working with large universes that lead to more statistically significant results than those in which "case studies" are used (Katz & Martin, 1997). Co-authorship reflects the formal result of the work

of a group of researchers. The analysis of its intensity involves the identification and mapping of scientific cooperation, at the micro (researchers), *meso* (institutions) or macro level (countries). In this context, Lancho-Barrantes et al. (2012) point out that although international collaboration tends to lead to greater visibility than national collaboration, the benefit of international collaborative activity may depend on the nations with which the activity is carried out.

At the international level, the share of a country's output consisting of first or last-authored international collaborations publications can be considered as strength of a science system (van Leeuwen, 2009). According to Moya-Anegón et al. (2013), countries with low scientific impact tend to benefit most by developing collaborative research with those who have, on average, a high impact. In this context, they hypothesize that the corresponding author has a special status in the team, in which he/she tends to play a leading role and denote him/her as the "research guarantor". They also consider that in negotiations within the team, obtaining the role of corresponding author is a manifestation of his/her importance and a reward for the additional efforts made in the group. In other words, the team delivering the corresponding author, whether at the macro (country), *meso* (institution) or micro (researcher) level, makes the greatest contribution to the research described in a paper and can thus be seen as the research guarantor. In order to validate this hypothesis, Moya-Anegón et al. (2013) evaluate at the macro level (countries) the distribution of the countries' scientific production, both in its entirety and the portion published as research guarantor, in order to analyse the correlation between the citation impact of these two publication sets.

In this context, corresponding authors - which are generally the first or the last author (Frandsen & Nicolaisen, 2010; Moya et al, 2013; Bordons et al, 2014; Hsiehchen, Espinoza & Hsieh, 2015) - confer greater acknowledgment, leadership, seniority or dominance; in contrast, absence in these roles could be associated with subordination or secondary role (González-Alcaide, Park, Huamaní & Ramos, 2017). Along these lines, Chinchilla-Rodríguez et al. (2018a) associate the corresponding authorship concept to the notion of scientific leadership. They analyse the leadership role in international collaborative research on the countries' scientific performance in biomedical research field, examining the several types of collaborative practices. They observed that internationalisation could be harnessed to improve specialization capacity, and scientific impact, integrating peripheral countries into a global network of scientific exchanges.

Corresponding authors are considered those taking responsibility on the activity coordination and the research communication process (ICMJE, 2017). In international collaborative research, this role is assigned to the main contributor of the research, and by extension, to his/her institutional affiliation. In this context, many studies have analysed the scientific role of countries and institutions as corresponding author concerning leadership when collaborating internationally. In an increasingly collaborative scientific structure, it is fair assuming correspondence author being a viable proxy of scientific leadership. In this context, analysing the leading role of each country in the scientific collaboration relationships provides an opportunity to deepen the understanding on the scientific system dependencies (van Leeuven, 2009; Bordons et al., 2014; Moya-Anegón et al, 2013; Chinchilla-Rodriguez et al, 2018a, 2018b).

Moya-Anegón et al (2018) focus on collaboration and on normalized impact, presenting an analysis of the relationship between collaboration (inter-national, national, and 'no collaboration') and field-normalized citation impact. They observe that the relationship between authorship and citation impact indicators is complex, reflecting a country's phase of scientific development and the coverage policy of the publication database. Moreover, they highlight that one should distinguish a genuine leadership effect from a purely statistical

effect in data analysis. In this context, they use the neutral term corresponding author instead of guarantor or leadership author.
In this research, as well as Moya-Anegón et al (2018) we use the term corresponding author.
In the current research, we aim to contribute to the understanding of the validity of measures based on corresponding authorship, evaluating their significance at the *meso* level, namely Brazilian institutions with at least 100 documents published in journals indexed by the multidisciplinary database Scopus from 2003 to 2015. More specifically, its goal is to analyse the influence of whether or not a Brazilian institution is the corresponding author on the scientific impact (citations) of its scientific documents, distinguishing between its collaborative articles with foreign institutions and those resulting from national collaboration. In this way the paper aims to examine whether this indicator aggregates valid and useful, additional information in the description of the Brazilian institutions' scientific performance.

Methodological procedures

We used a data set retrieved from the Scimago Institutions Rankings (SIR) portal, comprising a total of 607,454 papers published by Brazilian authors in Scopus in the period 2003-2005, without double count. The citable documents represents about 95% of this total and the 443 institutions with at least 100 citable documents accumulate more than 98% of the total output. The bibliometrics indicators for these Brazilian institutions were retrieved, of which 26% were in international collaboration, 33% were in national collaboration and 41% were with no inter-institutional collaboration. This way, the analysed international collaborative papers represent a lower proportion than national and non-collaborative collaborative papers.

For each institution, a series of indicators were recorded in an Excel Worksheet, listed in Table 1.

Table 1. Overview of bibliometric indicators used in the study.

Acronym	*Definition*
Publication-based indicators	
docc_ic_ca	Citable document number with international co-authorship in which it is the corresponding author
docc_ic_nca	Citable document number with international co-authorship in which it is not the corresponding author
docc_nc_ca	Citable document number with national co-authorship in which it is the corresponding author
docc_nc_nca	Citable document number with national co-authorship in which it is not the corresponding author
Citation impact indicators	
ni_ic_ca	Normalized impact for citable documents with international collaboration as corresponding author
ni_ic_nca	Normalized impact for citable documents with international collaboration as non- corresponding author
ni_nc_ca	Normalized impact for citable documents with national collaboration as corresponding author
ni_nc_nca	Normalized impact for citable documents with national collaboration as non-corresponding author

We define international co-authorship as co-authorship that involving authors from at least two different countries, and national co-authorship as that involving authors from at least two institutions from the same country and no authors from a foreign country. Papers published by authors from one single institution are considered not to involve institutional collaboration. This categorization is the same used for Moya-Anegón et al. (2018). However, here papers involving both national and international co-authorships are categorized as an international co-authorship (collaboration), while Moya-Anegón et al. (2018) assign such papers to both categories.

The indicators in Table 1 allow to examine the relationship between the condition of being the corresponding author (ca) in papers published both with national (nc) or international (ic) collaboration and their respective normalized citation impacts (ni).

To this end, first, we examined the four normalized citation impact indicators (ni_ic_ca, ni_ic_nca, ni_nc_ca, ni_nc_nca) concerning their dependence in relation to the scientific production volume of institutions.

Therefore, the 443 institutions were grouped into quartiles (Q1, Q2, Q3 and Q4), with respectively 110, 111, 112 and 110 institutions each one, in function on the citable document number (docc) produced by the institutions, in order to verify if any of these four impact indicators depends on the institutions' production volume.

Next, at a significance level of 0.01, ANOVA tests were applied to assess whether there is a statistically significant difference between quartiles concerning the four normalized citation impact indicators in Table 1.

Considering that, in this study, the ANOVA test applied to scientific production in international collaboration showed that citation impact intensity does not depend on the scientific document volume, for this type of collaboration the difference between ni_ic_ca and ni_ic_nca was evaluated considering the 443 institutions, by means of paired t-test, with a significance level of 0.01.

In the study of national scientific collaboration, the normalized citation impact was found to depend on the scientific production volume. Therefore, in this situation, we evaluated the difference between the normalized citation impact as corresponding author and that as non-corresponding author, per document production quartile. We used paired t-test with a significance level of 0.01.

Analysis of the results

Table 2 shows F and p values resulting from the ANOVA tests, in which the statistically significant difference between the impact indicators for citable document quartiles is identified by **. In this case, the normalized impact indicator depends on the scientific production volume of the institutions. Thus, in the case where there is a significant difference, the analysis of the impact indicator should be conducted for each quartile separately. In the other case, the normalized impact is analysed to the total of 443 institutions.

From Table 2, it can be observed that the analysis of normalized citation impact depends on the scientific production volume of the institutions when it involves national scientific collaboration, since in case where Brazilian institution is non-corresponding author (ni_nc_nca), there is significant difference between normalized impact in relation to the production quartiles.

Table 2. F and p values from ANOVA tests to normalized citation impact indicators comparing scientific production quartiles.

	Normalized citation impact indicators (ni)	*F*	*p*
National collaboration (nc)	ni_nc_ca	2.346	0.072
	ni_nc_nca**	5.802	0.001
International collaboration (ic)	ni_ic_ca	0.368	0.776
	ni_ic_nca	2.266	0.080

** Significance level = 0.01.

In the research developed with international collaboration, there are not statistically significant differences among the average citation impact per production quartiles, which may be due to the greater internal variability of the quartiles.

Influence of the corresponding author role on scientific impact in international scientific collaboration

Regarding the scientific production resulting from international collaboration (ni_ic), the impact does not depend on the production volume of the institution, that is, the impact does not depend on the production quartile to which the institution belongs (Table 2). Thus, in this scenario, it can be verified whether there is, overall, difference concerning the production citation impact between the status of corresponding author and non-corresponding author of the research, taken by institutions.
The t test showed that there is a statistically significant difference between the normalized citation impact as corresponding author (ni_ic_ca) and normalized citation impact as non-corresponding author (ni_ic_nca) when Brazilian institutions work in international collaboration (Table 3). In this situation when the institution does not take the corresponding author role the normalized citation impact average is equal to 1.48, whereas the normalized impact average is equal to 0.88 when the institution takes this role. This way when a Brazilian institution publishes in international partnership, it is strongly benefited if it does not take the role of corresponding author, since the difference between the normalized impact as non-corresponding author and as corresponding author tends to be equal 0.6021 to papers published in international co-authorship. Therefore, the benefit rate (or gain) in citation impact when a Brazilian institution does not act as corresponding author tends to be 40.6% higher than when it does, in papers published in international co-authorship.

Table 3. Comparison of impact indicators of 443 institutions' publications in international co-authorship, as corresponding author and non-corresponding author.

*ni_ic***	*Mean*	*Std. Deviation*	*CV*
ni_ic_ca	0.8794	0.74401	84.6%
ni_ic_nca	1.4815	1.20069	81.0%

**Paired t test: t = -9.472; p =0.000; significance level = 0.01.

Besides, when the Brazilian institutions are not the corresponding author in research published with international partners, they tend to reach normalized impact above the general average, that is, greater than 1. On the other hand, the research with international partners published by the Brazilian institutions as corresponding author get impact below the global average, although greater than those developed with national partners, as can be observed from values present in Table 5.

This result is aligned with that of Chinchilla et al (2018a) when analysing the benefits of international scientific collaboration by taking the scientific production of countries as aggregation level. In addition, the authors point out that this trend is distributed evenly in the various Brazilian institutions, especially among universities, where it is hypothesized that a strong expertise area may be overshadowed by areas of lesser scientific advancement. This way, disciplinary structure of Brazilian research can be conditioning the citation impact of their output. It has been argued that this factor accurately represents the strength of countries and by extension their institutions (Glänzel, 2001; López-Illescas, Moya-Anegón & Moed, 2011; Radosevic & Yoruk, 2014). The concentration or specialization in certain disciplines may have an impact on production and corresponding authorship, both nationally and internationally (Glänzel, Leta & Thijs, 2006).

With similar values to those present on Table 3, Chinchilla et al (2018b) observed the same tendency analysing the Nanoscience and Nanotechnology (NST) field, evidencing that Brazil has a normalized citation impact greater at non-leading position (non-corresponding author) on papers published in international collaboration. In this sense, these authors highlighted that international collaboration is the most determinant for high citation to the Brazil' scientific output in NST, since overall its papers in collaboration achieve normalized citation indexes above the world average. Besides, they observed that Brazil benefits greatly from collaboration with Switzerland at non-corresponding author position and obtains more impact when it takes corresponding author position in collaborative research with Argentina and Australia.

We clarify, of the 443 institutions analysed, 2 institutions with a low scientific production in the period (104 and 106 articles) did not have publications in international collaboration as corresponding author. For these institutions, ni_ic_ca was defined as being equal to zero. It is also clarified that only 9 other institutions had ni_ic_ca equal to zero, all belonging to the quartile of lesser scientific production (Q4), when published in international collaboration in the corresponding author position.

Besides, it should be noted that the six Brazilian universities that are among the top 500 institutions in the international rankings of research institutions, among them SIR (whose scientific production data are based on Scopus), are aligned with this general trend, because they presented highest normalized citation impact when they did not take the corresponding author position in their research published in international co-authorship, with the difference between ni_ic_nca and ni_ic_ca near to the aforementioned tendency; namely: University of São Paulo (USP), difference equal to 0.70; State University of Campinas (UNICAMP), difference equal to 0.59; São Paulo State University (UNESP), 0.57; Federal University of Rio de Janeiro (UFRJ), 0.69; Federal University of Minas Gerais (UFMG), 0.66; Federal University of Rio Grande do Sul (UFRGS), 0.56. It is noteworthy to highlight that, with the exception of UFRJ, all these institutions reached a normalized citation impact above 1 in the corresponding author position. As non-corresponding author, they achieved a normalized impact above 1.5. In addition, these institutions are the six most productive in the universe analysed, with production in the range of 28,000 to 118,300 documents and the percentage of scientific production published in international collaboration in the range of 20 to 30% of their respective total productions.

Of note are six outlier institutions with totally discrepant behaviour in relation to this trend, because they have a much greater citation impact at the role of corresponding author; they are: Presidente Carlos Carlos University, Amaral Carvalho Foundation, State University of Rio Grande do Sul, Center of Technological Education of Maranhão, Brazilian Army, National Institute of Science and Technology in Bioanalytics and Amazon Environmental Research Institute. It is important to mention that these institutions are predominantly specialized in few knowledge areas: Presidente Carlos University (Medicine, Agriculture and Veterinary);

Amaral Carvalho Foundation (Medicine); State University of Rio Grande do Sul (Physics, Materials Science and Agriculture), Federal Center of Technological Education of Maranhão (Physics, Engineering and Materials Sciences); Brazilian Army (Medicine and Engineering); National Institute of Science and Technology in Bioanalytics (Chemistry and Chemical Engineering); Amazon Environmental Research Institute (Environmental Sciences and Agriculture). In addition, it is significant to note that the Amazon Environmental Research Institute is the second institution with the highest percentage of documents published in international collaboration, corresponding to 94% of its production, especially with the USA. Besides, these outlier institutions published a small volume of scientific production (less than 240 documents), with the exception of Amazon Environmental Research Institute, with an amount of 3,115 published documents.

These results are in line with Glänzel, Leta & Thijs (2006) and Leta, Glänzel & Thijs (2006) that observed Brazilian science reached high relative impact values in Physics, Chemistry and Medicine fields, during 1991-2003. Also Grácio & Oliveira (2013) have pointed out that Brazilian research in Nursing and Environmental Sciences areas present high normalized impacts in relation to the world trend, during 1996-2011. Therefore, it is assumed that the high impact of the publications of these institutions as corresponding authors may be due to the acknowledged scientific background achieved by the Brazilian science in these knowledge fields at the scientific community.

On the other hand, there are 22 outlier institutions with behaviour that is totally distant from this trend, because they have a much greater impact in the status of non- corresponding authors. In general, these are non-university medical institutions, especially hospitals and other specialized medical institutes.

Influence of the corresponding author role on scientific impact in national scientific collaboration

Considering that the normalized impact of scientific production published in national collaboration (ni_nc) depends on the scientific production quartile to which the institution belongs (Table 2), that is, the impact depends on the scientific production amount, it was verified if there is a statistically significant difference on the normalized citation impact between both conditions: corresponding author and non-corresponding author, per document production quartile.

Table 4 presents the results of the paired t-test between ni_nc_ca and ni_nc_nca, by scientific production quartile, where it is observed that for quartiles Q1, Q2 and Q3 there is a statistically significant difference between the normalized impact for the both conditions: corresponding author and non- corresponding author. In the quartile Q4, where the scientific production amount is lower (between 101 and 184 published papers), no statistically significant difference was observed between ni_nc_ca and ni_nc_nca.

Table 4. Paired t-test to ni_nc_ca and ni_nc_nca, by scientific production quartile.

Difference between ni_nc_nca and between ni_nc_ca by quartile	*t*	*df*	*p*
Q1	-7.606	109	0.000**
Q2	-7.295	110	0.000**
Q3	-3.997	111	0.000**
Q4	-1.413	110	0.161

** Significance level = 0.01

From Table 5, the normalized citation impact of Brazilian institutions' publications that involves national collaboration was below the global average for both as corresponding author or as non- corresponding author. Besides, although for quartiles Q1, Q2 and Q3 it can be observed that institutions tend to benefit when they are non-corresponding authors (Table 4), this benefit is of minor importance. In publications involving national collaboration, the difference between normalized citation impact as non-corresponding author and as corresponding author not exceed an average increase of 0.085 (Q2), whereas in international collaboration, the difference between normalized impact as non-corresponding author and as corresponding author tends to be equal 0.6021, as previously observed.
It is highlighted that, on average, the works published in national co-authorship as non-corresponding author have a normalized impact equal to 0.66 for the most productive institutions (belonging to Q1) and equal to 0.57 for less productive institutions (belonging to Q4).

Table 5. Citation impact indicators' descriptive statistics for the publications with national collaboration concerning both conditions - corresponding author and non-corresponding author, by quartile.

Quartil	*Ni_nc*	*Mean*	*(ni_nc_ca - ni_nc_nca) mean*	*Std. Deviation*	*CV*
Q1**	ni_nc_ca	0.599	-0.057	0.109	18%
	ni_nc_nca	0.657		0.096	15%
Q2**	ni_nc_ca	0.532	-0.085	0.164	31%
	ni_nc_nca	0.617		0.145	24%
Q3**	ni_nc_ca	0.536	-0.081	0.226	42%
	ni_nc_nca	0.617		0.188	31%
Q4	ni_nc_ca	0.515	0.053	0.406	79%
	ni_nc_nca	0.568		0.187	33%

** Significance level = 0.01

Besides, in quartile Q1, institutions tend to achieve a normalized citation impact 0.057 higher at non-corresponding author condition than at corresponding author status. However, it is significant to note that taking into account only the 10 most productive institutions, this difference tends to be zero, since for them the average difference in the normalized citation impact as corresponding author and non-corresponding author is equal to -0.004; namely: USP, UNICAMP, UNESP, UFRGS, UFRJ, UFMG, Federal University of São Paulo (UNIFESP), Federal University of Santa Catarina (UFSC), Brazilian Agricultural Research Corporation (EMBRAPA) and Federal University of Paraná (UFRGS).
Furthermore, the observed trend for Q1 concerning national collaborative research tends to increase for the smaller scientific production quartiles; namely: for Q2, institutions tend to get 0.085 higher citation impact qua non-corresponding author; for Q3, institutions tend to reach 0.081 plus normalized impact as research non-corresponding author.
In the quartiles related to the largest scientific productions, Q1 and Q2, two outlier institutions have behaviour that differed significantly from the trends stand out because they had a greater impact in the corresponding author condition; namely: Rene Rachou Research Center (FIOCRUZ Minas Gerais) and Estácio de Sá University. It is important to mention that these institutions work predominantly in specific research areas, namely: Rene Rachou Research

Center in Medicine, Health, Biology and Agriculture areas; Universidade Estácio de Sá, in Biology area, especially Health and Medicine areas.
On the other hand, there are three outlier institutions concerning observed trends for Q1 and Q2 quartiles, because they have a larger impact at non-corresponding author condition. Of these, two institutions dedicate only to research, working in Engineering, Energy and Chemistry areas (Petrobrás do Brasil) and in Engineering and Computing areas (Renato Archer Information Technology Center). The other institution is a university (Federal University of the South Frontier) that focus its research in Agriculture and Biochemistry areas.
It is noted that in the quartiles with the lowest scientific production institutions (Q3 and Q4) there is a greater presence of outlier institutions, resulting from the high variability detected for these quartiles, as it was observed in the Coefficient of Variation presented in Table 5.

Final considerations

This research aimed to examine whether or not the normalized citation impact is influenced by the condition of being corresponding author, for the Brazilian institutions that carried out research in scientific collaboration. In this context, this question was analysed in two different collaboration perspectives - international and national - in order to analyse if a the status of corresponding author adds additional information in the scientific performance description of the Brazilian institutions.
In international collaboration, it was observed that the normalized citation impact achieved by Brazilian institutions tends to depend upon the corresponding author status and that under the non-corresponding author condition, they tend to have a very significant benefit when the paper has as its corresponding author a foreign institution, independent of the published paper quantity or of the institution's consolidation in international rankings of academic-scientific performance. This result suggests that Brazilian institutions have sought to associate themselves with international institutions of recognized academic and scientific primacy.
Besides, the results obtained in this research align with those found by Chinchilla et al. (2018a), which show that, with the exception of the United States, Canada and some European countries (e.g., Sweden, France and the Czech Republic), which have consolidated research systems and the lesser differences in all types of authorship, non-leading internationally collaborative articles reach higher values than leading international and domestic papers, in this order. That means that publishing with internationally leading partner's benefits and strengths the lesser scientific highlight countries (González Alcaide et al, 2017). That could mean that Brazilian institutions benefit of the leadership of foreign institutions to improve the absorptive capacity, visibility and integration into global collaboration research networks (Chinchilla et al, 2018a).
In national collaboration, it was observed that the normalized citation impact tends to be associated with the scientific production quantity published by the institution, with larger institutions tending to reach higher citation impact than those with smaller scientific production. Besides, institutions with smaller scientific production presented more dispersed performance than those most productive. Thus, the role of corresponding author was analysed in function of the amount of scientific production.
It was noted that also in national collaboration, the institutions benefit as non-corresponding author, although the most productive institutions reach a more significant normalized citation impact, regardless of whether they take the corresponding author status or not. However, we considered that this difference, although statistically significant, does not bring enough convincing evidence to influence the practice of the institutions' scientific policies, since the

value added in the normalized citation impact by the non-corresponding author condition is not sufficiently significant to influence the scientific decision-making.
Thus, the indicator of corresponding author provides additional information relevant to Brazilian institutions in international collaboration, but not in the case of national institutional co-authorship, which is more influenced by the institution's recognized scientific tradition and publishing strategies/practices.
Therefore, in summary, we find that in the case of Brazil, the differences in citation impact between Brazilian and foreign institutions are generally much larger than those between Brazilian institutions. In this sense, in national collaboration, corresponding authorship hardly affects citation impact. It is hypothesized that Brazilian institutions have similar and low significant international scientific impact. These are questions to be thought and researched further in follow-up studies.
The results obtained in this research arouse several opportunities for future research, some of which are presented below.
In relation to scientific production in international collaboration, the applied statistical test showed that citation impact intensity does not dependent on the scientific output in terms of number of published documents. However, in general, scientific literature has showed that size matters in international collaboration: the greater capacity of institutions (more investment, infrastructure, training skilled researchers, among other), the more internalised their production (number of papers in non-institutional collaboration and national collaboration). On the other hand, small institutions tend to depend more on international collaboration for their output than big institutions, in terms of the number of papers. Besides, international collaboration has a higher citation gain than national or intra-institutional collaboration, especially in the case of emergent and developing countries, such as Brazil.
Hence, we consider that more in-deep analysis in future research should be developed in order to understand this result obtained in this research.
Besides, it is considered that this approach can contribute to an understanding of how national or international scientific output could be improved harnessing partner' specialization strengths. In this sense, it is also suggested to analyze more granular differences in international collaborative dynamics (e.g., distinguishing between regional and extra-regional collaborations) and across scientific disciplines or fields, considering that can help to identify competitive scientific fields at the national and international level and those that need being fortified by international partnership.
Another suggestion for furthering this research is the construction of visual representations which may contribute to identify the collaborator countries involved in this production, in the sense of encourage mobility programs for reinforce their absorptive capacities as well as providing the understanding the knowledge flow in international research dynamics and signalising the specialization interrelationships, their intensities that reverberate on scientific citation impact. This can provide methodological subsidy to decision makers on the design of countries' strategic plans for science and technology.
Another issue that arises in this research relates to the language of publication. In this sense, it is suggested that research is developed analysing the influence of the language of publication on citation impact. Along this line, for several non-English speaking countries, the inclusion of a higher number of national journals is associated with a drop in citation rates, because, among other factors, the fact of most journals are not published in English. A high concentration of papers published in Brazilian journals means a constraint concerning the potential number of readers to a Portuguese language research community.
Following the idea of the importance of the venue of publication, we can hypothesize that the difference in citation impact, acting as corresponding author or not in international collaboration, could be associated with the fact that collaborators (acting as corresponding

authors) are publishing in better journals (in terms of journal impact measures such as SJR) than Brazilian researchers do in the same role. That should be checked in future research in order to analyse the prestige of journals where Brazilian institutions publish in collaboration with international or national partners contributing to constructive reflection upon publishing and editorial practices (Chinchilla, Miguel & Moya-Anegón, 2015).

The evaluation regarding the involvement of international and national funding organizations is significant to determine the citation impact of publications. The higher collaborative nature of funded publications may play an important role in setting and facilitating collaborations among scholars and countries (Costas & Van Leeuwen, 2012; Defazio et al., 2009). In terms of citation impact, funded publications exhibit a higher average field-normalized impact as compared to those publications that do not mention any funding source (Kozma, Calero-Medina & Costas, 2018). This can lead to interesting perspectives for future research concerning the analysis of funding acknowledgements embedded in publications in relation to corresponding authorship, i.e., if the country of the corresponding author is the same that the funding agency.

As a last suggestion, it is considered relevant to evaluate whether the corresponding author makes a greater contribution than the other authors in by-line and, in this sense, can be considered the research guarantor. In this context, a qualitative validation study of the indicator of corresponding author is also suggested, through the analysis of the Brazilian authors' understanding involved in papers published in co-authorship.

Finally, it is concluded that the indicator of corresponding author brings significant contributions to the understanding of the Brazilian institutions' scientific impact reach, by enlarging the understanding the influence of the corresponding author status, or not, on the recognition of their science produced by the scientific community and by the reward system of science. It also contributes to bibliometrics studies for describing characteristics of scientific production not yet identified by the other usual indicators.

References

Alonso Arroyo, A., de Oliveira, E. F. T., Grácio, M. C. C., Pandiella, A. & Aleixandre Benavent, R. (2016). Un análisis bibliométrico en el área de la Medicina: colaboración científica entre Brasil y España (2002-2011). *Investigación Bibliotecológica*, 30(69), 205-229.

Bordons, M., González-Albo, B., Aparicio, J. & Moreno, L. (2014). The influence of R&D intensity of countries on the impact of international collaborative research: evidence from Spain. *Scientometrics*, 102 (2), 1385-1400.

Chinchilla-Rodríguez, Z., Miguel, S. & Moya-Anegón, F. (2015). What factors are affecting the visibility of Argentinean publications in human and social sciences in Scopus? Some evidences beyond the geographic realm of the research. *Scientometrics*, 102 (1), 789-810.

Chinchilla-Rodríguez, Z., Larivière, V., Costas, R., Robinson Garcia, N. & Sugimoto, C.R. (2018a). Building ties across countries: international collaboration, field specialization, and global leadership. In *Proceedings of the 23rd International Conference on Science and Technology Indicators*. Leiden: Centre for Science and Technology Studies (CWTS). From: https://openaccess.leidenuniv.nl/handle/1887/65228.

Chinchilla-Rodríguez, Z., Miguel, S., Perianez Rodríguez, A. & Sugimoto, C.R. (2018b). Dependencies and autonomy in research performance: examining nanoscience and nanotechnology in emerging countries. *Scientometrics*, 115 (3), 1485-1504.

Costas, R. & Van Leeuwen, T. N. (2012). Approaching the 'reward triangle': General analysis of the presence of funding acknowledgments and 'peer interactive communication' in scientific publications. *Journal of the American Society for Information Science and Technology*, 63(8), 1647–1661.

Defazio, D., Lockett, A. & Wright, M. (2009). Funding incentives, collaborative dynamics and scientific productivity: Evidence from the EU framework program. *Research Policy*, 38(2), 293–305.

Frandsen, T.F, Nicolaisen, J. (2010). What is in a name? Credit assignment practices in different disciplines. J Informetrics, 4 (4), 608–617.

Glänzel, W. (2001). National characteristics in international scientific co-authorship relations. *Scientometrics*, 51, 69–115.

Glänzel, W. (2003). *Bibliometrics as a research field*: a course on theory and application of bibliometric indicators. Belgium: Katholieke Universiteit Leuven.

Glänzel, W. & Lange, C. A. (2002). Distributional approach to multinationality measures of international scientific collaboration. *Scientometrics*, 54 (1), 75-89.

Glänzel, W., Leta, J. & Thijs, B. (2006). Science in Brazil. Part 1: A macro-level comparative study. *Scientometrics,* 67 (1), 67-86.

González-Alcaide G, Park J, Huamaní C. & Ramos, JM. (2017). Dominance and leadership in research activities: Collaboration between countries of differing human development is reflected through authorship order and designation as corresponding authors in scientific publications. *PLoS ONE*, 12(8), e0182513.

Grácio, M.C.C. & Oliveira, E. F. T. (2013). Normalized Indicators of the International Brazilian Research: a scientometric study of the period between 1996 and 2011. In J. Gorraiz, E. Schiebel, C. Gumpenberger, M. Hörlesberger & H. Moed (Eds.), *Proceedings of ISSI 2013 - 14th International Society of Scientometrics and Informetrics Conference* (pp. 2069-2071). Vienna: Facultas Verlags- und Buchhandels AG.

Hsiehchen, D., Espinoza, M. & Hsieh, A. (2015). Multinational teams and diseconomies of scale in collaborative research. *Science Advances*, 1 (8), e1500211.

ICJME (2017). Recommendations for the Conduct, Reporting, Editing, and Publication of Scholarly Work in Medical Journals. From: http://www.icmje.org/icmje-recommendations.pdf

Katz, J. S. & Martin, B. R. (1997). What is research collaboration? *Research Policy*, 26, 1-18.

Kozma, C., Calero-Medina, C. & Costas, R. (2018). Research funding landscapes in Africa. In C. Beaudry, J. Mouton & H. Prozesky (Eds.), *The Next Generation of Scientists in Africa* (pp. 26-42). Cape Town: African Minds.

Lancho-Barrantes, B.S., Guerrero-Bote, V. P., Chinchilla-Rodríguez, Z. & Moya-Anegón, F. (2012). Citation Flows in the Zones of Influence of Scientific Collaborations. *Journal of the American Society for Information Science and Technology*, 63, 481-489.

Leta, J., Glänzel, W. & Thjis, B. (2006). Science in Brazil. Part 2: Sectoral and institutional research profiles. *Scientometrics,* 67 (1), 87–105.

López-Illescas, C., de Moya-Anegón, F. & Moed, H.F. (2011) A ranking of universities should account for differences in their disciplinary specialization. *Scientometrics*, 88 (2), 563-574.

Moya-Anegón, F., Guerrero-Bote, V.P., Bornmann, L. & Moed, H. F. (2013). The research guarantors of scientific papers and the output counting: a promising new approach. *Scientometrics***,** 97, 421–434.

Moya-Anegón, F., Guerrero-Bote, V.P., López Illescas, C. & Moed, H. F. (2018). Statistical relationships between corresponding authorship, international co-authorship and citation impact of national research systems. *Journal of Informetrics*, 12, 1251-1262.

Persson, O., Glänzel, W. & Dannell, R. (2004). Inflationary bibliometric values: the role of scientific collaboration and the need for relative indicators in evaluative studies. *Scientometrics*, 60 (3), 421-432.

Radosevic, S &Yoruk E. (2014). Are there global shifts in the world science base? Analyzing the catching up and falling behind of world regions. *Scientometrics*, 101(3), 1897-1924.

Van Leeuwen, T.N. (2009). Strength and weakness of national science systems: A bibliometric analysis through cooperation patterns. *Scientometrics*, 79 (2), 389-408.

Designing healthy and sustainable food systems: how is research contributing?

Agénor Lahatte[1], Élisabeth de Turckheim[1,2] and Lucile Chalumeau[1]

[1] agenor.lahatte, elisabeth.de-turckheim, lucile.chalumeau)@hceres.fr
OST, High Council for Evaluation of Research and Higher Education (Hcéres), 2 rue Albert Einstein, 75013, Paris (France)

[2] Délégation à l'évaluation, INRA, 147 rue de l'Université, 75338 Paris Cedex 07 (France)

Abstract

A method for delineating the research area addressing a complex global problem is developed in a context where no core set of documents is available and where the objectives of the research are described as the production of broad knowledge related to societal outcomes. The method, entirely based on textual analysis has two steps: (a) multi-terms selected from a policy document are used to retrieve documents from a publication database, (b) a LDA topic model fitted to the retrieved corpus allows to identify irrelevant topics and to remove the corresponding documents. Once the corpus is cleaned, topics are clustered on the base of their semantic proximities and of their co-occurrence in documents for structuring the domain into main research themes. The method is applied to a selected facet of the food security challenge and provides a representative sample of research works in the domain. Structuring the corpus into six research themes allows to compare the thematic profiles and the research practices of twelve countries most involved in the domain.
Keywords: corpus delineation, societal challenge, topic model, food security, sustainable food system

Introduction

Assessing the contribution of research addressing complex global challenges has become increasingly important in science policy as research framework programme or strategies tend to align on societal demands. This raises the issue of identifying the research areas that may contribute to the challenge either to define scientific programmes or to assess the achievement of such research policies. In these situations, the first problem is to delineate a relevant corpus of scientific documents.

This issue of delineating a research domain is regularly addressed in science policy or scientometrics. This is for instance a key point for emerging scientific domains (Milanez, Noyons & de Faria, 2016; Raimbault, Cointet & Joly, 2016). In such cases, usual methods select specific scientific multi-terms and, after some trial and error, use these multi-terms as queries to retrieve documents from a database. The resulting corpus is then either cleaned or extended to reach a satisfactory balance between recall and precision of the final corpus.

However, when the objectives of a programme are oriented towards social needs, they are not directly characterized by scientific keywords as they point to expected societal outcomes of research and should leave open a diversity of scientific options (Wallace & Rafols, 2015). In such cases, queries have to be more general to retrieve documents in any possibly contributing field. But queries based on general language terms, partly because of words polysemy, may harvest documents that are out of the scope of the programme. Therefore an intensive cleaning step may be necessary.

Topic models are adapted for screening large datasets. They only rely on textual information and have been presented as an efficient method to extract relevant documents from very large corpora (Klavans and Boyack, 2014) or to find the thematic structure of sets of non academic documents as web pages or press articles (Di Maggio, Nag and Blei, 2013). We show here

how topic models can be used to clean a corpus that may contain important subsets of irrelevant documents.

As a case study, we choose to delineate a corpus of publications addressing the challenge of food security and we focus on a facet of the challenge entitled *Designing safe, sustainable and innovative food systems* from a policy document of the French Ministry of Higher Education and Research (MESRI, 2014). We extract from this document the main questions addressed to researchers and delineate a related corpus. This corpus is then analysed through six main research themes. We compare the scientific production of various countries to the domain and characterize the degree of interdisciplinarity of themes and of countries contributions.

Corpus delineation

The delineation based on a policy document consists in two steps: first a selection of key phrases that are used as queries to retrieve documents from the chosen database and, second, a cleaning step of the retrieved corpus.

Choosing queries

The policy document is a workshop report of the Expert group of the Ministry (MESRI, 2014) that focuses on a set of selected research questions which are: How to improve the efficiency of the food supply chain? What are the determinants of consumers' behaviour? How and why diets are changing? How diets are influencing people nutritional status and health? What are the environmental and social impacts of food systems? Which policy and institutional actions could enhance food security and nutrition?

We reorganize these research questions as 8 items. An overall issue and 7 elements of food systems according to the conceptual framework of food systems of the FAO Committee on World Food Security (HLPE, 2017, page 26)

1. Sustainable food systems ensuring food security
2. Food chain efficiency with a focus on food process optimisation, food circulation, localized agri-food, food losses and waste reduction
3. Food environments impacting consumers' behaviour: food availability, food information...
4. Consumers' behaviour
5. Diets and diet transitions
6. Environmental and social impacts of food systems
7. Nutrition and health outcomes of diet patterns
8. Policy and institutional actions to move towards sustainable food systems.

According to these research questions, we design 32 queries (Table 1).

We use the OST home version of the WoS database restricted to citable documents (article, letter, review), with at least one address and a WoS category. We query for documents published between 2012 and 2017 where the selected key phrases appear in the text that merges title, abstract and authors' keywords of each article.

Table 1. Selected research questions and associated queries

Questions to scientific research Item #	Query#	Queries
1 Sustainable and healthy food systems, that ensure food and nutrition security	1	sustainable AND food system%
	2	sustainable food
	3	food system% AND (safe OR safety)
	4	food system% AND (secure OR security)
	5	food security OR nutrition security
2 Food chain efficiency: Optimisation of food processing and production, optimization of food circuits, reduction and recycling of food losses and waste, localized agri-food production	6	(efficien% OR optim%) AND (food process% OR food production)
	7	eco\-design AND food
	8	food preservation AND (technolog% OR process%)
	9	preservation AND (food process% OR food technol%)
	10	optim% AND (food circuit% OR food supply)
	11	short AND (food circuit% OR food supply)
	12	food waste%
	13	food loss%
	14	local% (agrifood OR agri\-food OR agrofood OR agro\-food)
3 Food environments that impact consumer behaviour: Food availability and supply, food information, food quality and safety	15	food availab% and (choice% OR consumer%)
	16	food supply AND (choice% OR consumer%)
	17	food (information OR advertising OR marketing OR promotion) AND (choice% OR consumer%)
	18	food quality AND (choice% OR consumer%)
	19	(food safety OR safe food) AND (choice% OR consumer%)
	20	(healthy food OR healthy eating) AND (choice% OR consumer%)
4 Consumer behaviour	21	consumer% behavio% AND food
	22	food choice% AND consumer%
5 Diets and diet transitions	23	(food OR diet OR dietary) pattern%
	24	(food OR diet OR dietary OR nutrition) transition%
6 Impacts of food systems	25	social impact% AND (diet OR dietary OR food system%)
	26	environment% impact% AND (diet OR dietary OR food system%)
	27	impact% on environment AND (diet OR dietary OR food system%)
7 Nutrition and health outcomes	28	health benefit% AND (diet OR dietary OR food pattern%)
	29	healthy (diet OR dietary OR food pattern%)
	30	health risk% AND (diet OR dietary OR food pattern%)
8 Policy and governance	31	(food OR nutrition) polic%
	32	(food OR nutrition) governance

Cleaning the initial corpus by identifying irrelevant topics

The resulting corpus C1 contains 24,890 documents. To screen this corpus and control that there are not too many documents out of the scope of the study, we fit a LDA topic model (Blei, Nag & Jordan, 2003; Blei, 2011) to identify possible irrelevant documents. We choose the grain to examine the corpus as approximately 1,000 documents so that a model with 25 topics is used to scrutinize corpus C1. Among the 25 fitted topics, 3 of them are considered out of the scope of the study. A topic on ecology, with papers on marine ecology, evolutionary developmental ecology (evo-devo) and other specialities in ecology was collected with the queries on diet patterns or on food availability, because these terms are common for human diets and for other animal species. A second topic on plant genetics and plant science is considered marginal for this study, despite the fact that improving crop yield with selection

and genetic engineering is a way to improve food production. We consider that the articles collected here do not correctly represent the whole research on crop genetics and plant physiology. We therefore decide not to include this research area in our study. The third topic is focused on the impact of particular food components on the metabolism, through clinical and experimental research. Again we decide not to include this level of scientific investigation in the study.

As the same queries retrieve documents in both relevant and irrelevant topics, it is not possible to remove initial queries. It also appears difficult to refine queries, for instance so that they only apply to human diets. Therefore, in order to remove the documents focused on unwanted topics, we use a similar method as Milanez, Noyons & de Faria (2016) and select terms that are specific to these irrelevant topics to remove off-domain documents. Such terms are easily found with the LDAvis software (Sievert & Shirley, 2014) setting the parameter λ of the *relevance indicator* very low, thus selecting terms with a high probability in the topic and a very low probability in the rest of the corpus. These specific terms are used as counter-queries: documents of corpus C1 that use them are removed to get corpus C2. We also remove articles in journals of WoS categories of the specific scientific approaches that we do not want to include in our study (Table 2). This provides a second corpus C2 of 20,500 documents where the irrelevant topics do not appear any more.

Table 2. Counter-queries and selected WoS categories to remove documents focused on irrelevant topics

Irrelevant topics	Counter-queries	Removed WoS categories
Ecology, marine and freshwater biology, evo-devo	prey%, predator%, trophic, foraging, benthic, coral, reef%, juvenile, zooplancton, phytoplancton, pelagic, invertebrate, neolithic, archeological, webs, nest%, catches	ECOLOGY MARINE & FRESHWATER BIOLOGY ZOOLOGY ENTOMOLOGY ORNITHOLOGY
Plant genetics, plant science	genome%, landraces, accession%, loci, allele%, allelic, qtl%, arabidopsis, transcriptom%, snps, nucleotide, chromosome%, microsatellite%, heterosis, heterozygo%, inbred, barcoding	PLANT SCIENCES GENETICS & HEREDITY
Metabolism and endocrinology	insulin, rats, mice, hdl (high density lipoprotein), ldl (low dentity lipoprotein), lipoprotein, nafld (non alcoholic fat liver disease), hfd (high fat diet), crp, (C reactive protein), adipose, adinopectin, leptin, tnf (tumor necrosis factor), homa (homeostasis model assessment), steatosis, interleukin, lps (lipopolysaccharide), pcos (polycystic ovary syndrome), wistar (a laboratory rat), ppar (peroxisome proliferator-activated receptor), cortisol, macrophage%	ENDOCRINOLOGY & METABOLISM

Semantic analysis of the corpus

Fitting a LDA model with 20 topics on this C2 corpus shows that the chosen research questions in Table 1 are present in this corpus and that one to three topics correspond to each question. Table 3 displays a title of these topics based on the most relevant terms (Sivert & Shirley, 2014) and on the titles of the 30 articles most focused on each selected topic. We

organize (and renumber) them into 8 groups according to the initial research questions. Some questions of Table 1 are merged in the same topic - as the issues of food policies and systems governance that are merged into the first group. In Table 3, the fist column shows the relationship between the topics and the initial questions.

Table 3. The 20 topics of corpus C2

Item #	Topic #	Topic title
1,8	1	Food systems: sustainability, policies, governance
	2	Food security: markets and trade, agriculture efficiency, land management
	3	Food insecurity: social factors
2	4	Food processing technologies
	5	Food waste treatment and biogas production
	6	Detection and control of food microbial contamination
3,4	7	Factors impacting consumers' behaviour
	8	Nutritional education
	9	Factors influencing young people's nutrition
5	10	Diet patterns
6	11	Environmental impact of the food supply chain
7	12	Health risks associated with diet pattern, health benefits of (mediterranean) diet
	13	Health risks: children and adolescents lifestyle and obesity
	14	Health risks associated with diet pattern: diabetes, depression, cancer...
	15	Food contaminants and dietary exposure to heavy metals, pesticides residues...
7	16	Healthy food components: nutraceuticals, probiotics, prebiotics....
	17	Healthy food components: antioxidants (unsaturated fatty acids, phenolic components...)
	18	Agriculture and climate change
	19	Farm animal feeding
	20	Methods and models for estimation of crop production and monitoring

The first group is composed of articles on food and nutrition security and policies (Topic 1) and on economic and trade aspects of food systems (Topic 2 and Topic 3). Three topics are focused on food technology: food processing (Topic 4), treatment of food waste and manure for biogas production (Topic 5), detection and control of food microbial pathogens, food maturation and preservation with food microbiota (Topic 6). Topics 7, 8 and 9 are devoted to consumers' behaviour and its determinant factors as food quality perception, food marketing and nutritional education. Topic 10 is a general topic on diets and diet changes. Health risks of diets are present in Topics 12, 13 and 14 with Topic 13 focused on young people diet and lifestyle. Food contaminations by heavy metals and pesticides are explored in Topic 15. One topic is about environmental impacts of food systems (Topic 11). Finally, Topics 16 and 17 are focused on beneficial food components as antioxidants and nutraceuticals (food with health benefits beyond their nutritional value). As shown on Figure 1, topics in the same topic group have similar vocabularies.

Three last topics appear in this corpus though there were not explicitly requested: Topic 18 that deals with the interactions between agriculture and climate change is linked to food (in)security and therefore close to Topics 1 and 2. It also has a vocabulary similar to the topic on environmental impacts of agri-food systems (Topic 11). However, the retrieved publications do not represent properly the whole research on climate change and agriculture interactions but only those articles that explicitly cite the impact on food security. This domain related with the food security issue should either be more intensively included or treated separately. Topic 20 is an ancillary topic, with methodologies for data collection and analysis and tools for image treatment, surveys etc. Finally, Topic 19 on farm animal feeding (mainly zootechnics) is out of the scope of the study. This topic was not detected in the first topic model and we later remove the documents mainly focused on this topic (326 documents).

The 17 first topics properly cover the expected research themes, except the social impacts of food systems that do not appear as a specific topic. This research question is not visible at this grain, probably merged in Topic 1.

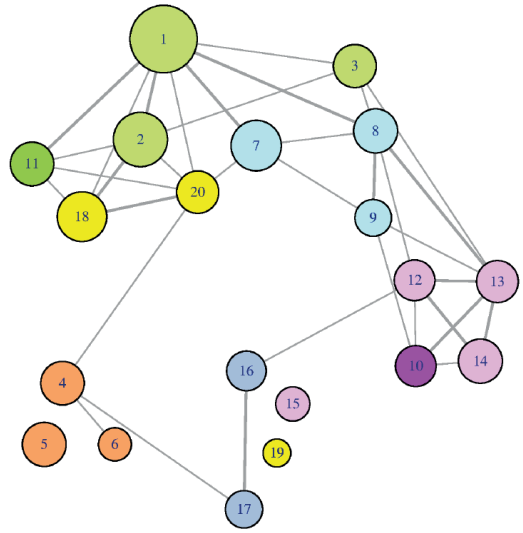


Figure 1. Map of the 20 topics of corpus C2. Topic positions are related to topic vocabulary: two topics are close in the figure if they use similar vocabularies. Bubble surface is proportional to topic weight. Edges show two levels of co-occurrence of topics in documents with a threshold chosen so as to select about 20% of the total number of possible edges (Jensen-Shannon divergence less than 0.5 or less than 0.55) .

Topic co-occurrences

In a topic model, documents combine topics. Various types of topic sharing may occur. At one end, a topic could appear mainly in specific documents where the topic weight is very high

and no other topic is important. At the other end, an ancillary topic appears in many documents with a moderate weight. This is often the case with topics gathering general scientific terms as jargon or about methodology for data collection and analysis as statistics, data modelling, image processing, survey methodology etc. that are often associated with core topics. Between these two extremes types, some topics could play both roles and groups of related topics could compose small cliques. Various measure of topic co-occurrence are possible. In a preceding paper (Cassi, Lahatte & Rafols, 2017), we used the number of documents where the weight of each of two topics is above a threshold (say 25%) to measure topic co-occurrence. In this work, we use a dissimilarity measure between the two vectors of the document:topic matrix (in this case, the Jensen-Shannon divergence).

Figure 1 shows that Topic 1 is the most connected topic with 7 edges (or degree 11 if a weight 2 is given to the 13 strongest edges of the network). Not surprisingly, the methodology Topic 20 is sharing numerous documents with other topics (6 edges) as well as Topic 8 on nutritional education (6 edges). It makes sense to say that nutritional education is a factor driving consumers' behaviour (Topics 7 and 9) and is related with social aspects of food systems (Topic 3) and with food policies (Topic 1). Topic 13 on young people diet (6 edges) shares documents with Topics 8 and 9 (on young people exposure to harmful influences) and with papers on health risks of inappropriate diets. At the other end, topics on food components (Topics 15, 16, 17) are not sharing many documents with the rest of the corpus. The three topics on food technology are also mostly specialised, with only two edges with other topics.

This connected network suggests to reconfigure the initial research questions of the policy document into five groups of topics or five research themes structuring the domain. In addition, considering that the nutritional and health status of young people is an important societal challenge and that topics 8, 9 and 13 are strongly linked, we also select a sixth theme, partially overlapping two other themes (Table 4). We leave aside four marginal topics thus focusing on a third corpus C3.

Table 4. The six reconfigured research themes

	Item #	Research theme	Topics	Document #	Theme %
Th1	1,6,8	Food security, food systems, social factors, policies	1, 2, 3, 11	6257	0.31
Th2	3,4	Consumers' behaviour and factors impacting it	7, 8, 9	3159	0.15
Th3	5,7	Diet patterns and health risks	10, 12, 13, 14	3629	0.18
Th4	7	Healthy foods	16, 17	1563	0.08
Th5	2	Food chain efficiency	4, 5, 6	2874	0.14
		Marginal topics	15, 18, 19, 20	3018	0.15
Th6		Young people nutritional and health status	8, 9, 13	2572	0.13

Analysis of country publications

Country publications

Among the countries with the largest contribution to the corpus C3, USA, Great Britain and China have more than 1,500 publications in the domain during the period 2012-2017 (Figure 2, whole counting). Five other European countries (Italy, Spain, Germany, The Netherlands and France) have more than 600 publications. Outside Europe, Australia, Canada, Brazil and India are also among the 12 most most productive countries. We focus on these 12 countries to compare their thematic profiles.

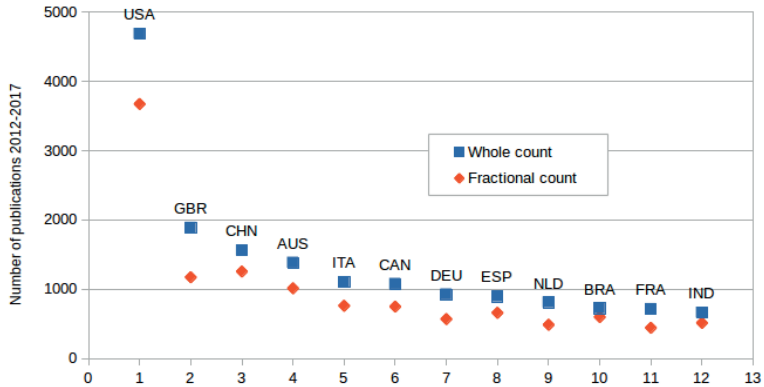


Figure 2. Number of publications of the 12 countries with the largest contributions to corpus C3

Country thematic profiles

The thematic profile of a country is obtained by attributing each document to its heaviest topic. This is a more convenient rule than assigning fractions of documents as the underlying model assumes. We also assign a document to a country as soon as it has an address in the country (whole counting).

Thematic profiles of countries show major differences (Figure 3). Canada, Great Britain, Germany and The Netherlands have a higher involvement than the world average on food systems sustainability and food policies while Spain is much less covering this theme than the world average. China and India are more involved in improving food chain efficiency but are less specialised on diet health risks. In an opposite way USA, Australia, and The Netherlands have a higher relative commitment on food risks, balanced with a low one on food chain efficiency. India, Italy and Spain seem to publish more research about healthy food, possibly in relation with regional diets as the Mediterranean or vegetarian diet. Two countries, USA and Australia, are more concerned by young people diet and lifestyle than the rest of the world and this may be related with the high prevalence of obesity in USA, Canada and Australia (OECD, 2017). These first findings seem consistent with the general context of food patterns and food system issues in the different countries.

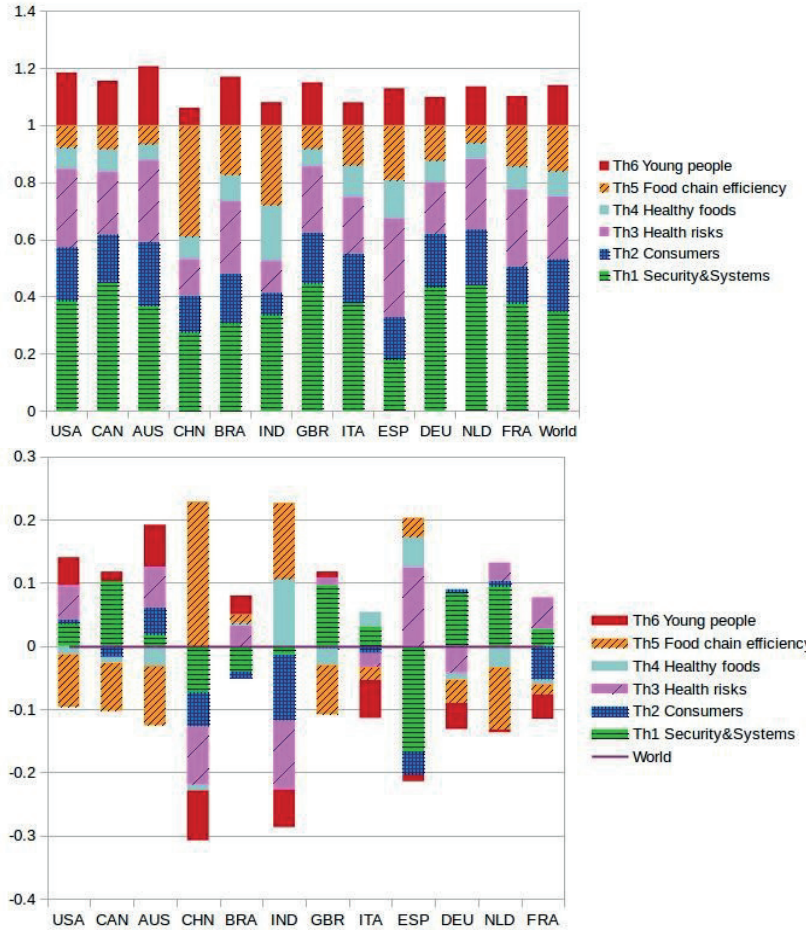


Figure 3. Thematic profile of 12 countries: distribution of publications of corpus C3 by country (whole counts) in each thematic group (top figure), difference of proportions with the world (bottom figure)

How interdisciplinary are themes and country contributions?

An expected feature of the research oriented by a complex global problem is its ability to cross disciplinary approaches (Ledford, 2015; Molas-Gallart, Rafols & Tang, 2014). It is not only the disciplinary diversity of the articles that is relevant but also the ability to combine different approaches, theories, concepts or methods in a same research work. The repartition of the corpus in WoS categories confirms the diversity of scientific fields involved but it does not inform about interdisciplinary that is achieved at the article level. To measure it, we use the integration score proposed by Porter, Cohen & Roessner (2007) based on the Rao-Stirling diversity index of the categories of references and we compare this indicator between the 6 themes and the 12 countries.

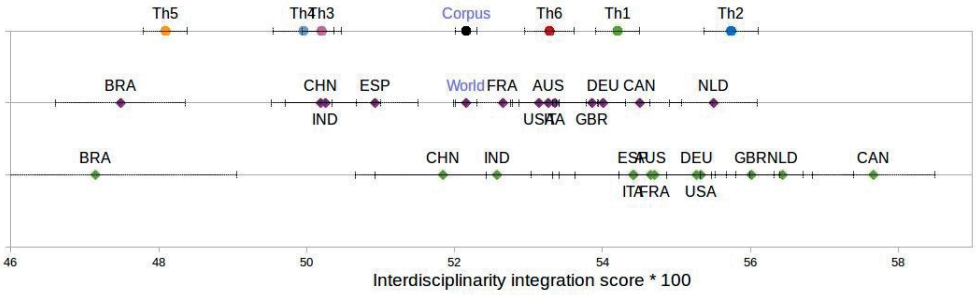


Figure 4. Interdisciplinary integration scores of countries, of themes and of countries for Theme 1, with90% confidence intervals

The most interdisciplinary theme is Theme 2 on consumers behaviour. Theme 1 on food systems and policies and Theme 6 on young people diet are also more interdisciplinary than the three other themes. More data on the position of country indexes by theme shows that Brazil has very low interdisciplinary indexes in themes 1, 2, 4 and that India has the lowest index in Theme 3, just before Brazil (Figure not displayed). As China, India and Spain are most specialised in the less interdisciplinary themes, this also contributes to their low overall interdisciplinarity index. At the other side, The Netherlands have the highest index in five themes. For Theme 1, the striking position of Canada is to be mentioned, significantly above every other country ans not statistically different from The Netherlands (Figure 4). The three most interdisciplinary countries for this theme (Canada, The Netherlands, Great Britain) are also the three countries most involved in the theme. This suggests a correlation between the specialisation in this theme and the disciplinary integration of these studies.

Discussion

The first results about the country profiles on the six final research themes suggest that the delineation method is relevant to select an appropriate corpus of publications representing the research on food security and food system sustainability. However, some methodological issues remain insufficiently resolved. The query selection is a trial and error process with examining samples of retrieved documents to further improve the queries. The part of arbitrary decision of this step is reasonably corrected by the cleaning process.

The cleaning process itself, based on a topic model, is dependent on various parameters. The number of topics is the first issue and this number is related with the desired degree of precision of the process. With a lazy choice (an average of 1,000 documents per topic), we had to remove a smaller off-domain topic at the second step. The second issue is that the fitting algorithm has a part of randomness and with another algorithm or just with another seed, topics may be different. As main topics are generally found in each fit with the same core terms, small topics may be arranged differently. However, the cleaning step is finally defined by the selected counter-queries and rejected categories. Though they have been

suggested by topics, they are to be validated *per se* and this is not too much time consuming for experts.

These points are to be taken into consideration but, in our experience, they are manageable and topic modelling appears to be an interesting method to assist and correct a delineation task in a context of imprecise queries.

Two other issues have not been considered in this work and should be solved before the method is used in an assessment process. First, the cleaning process makes it possible to control the precision of the delineation process, but there is no control of the recall so far. Second, the choice of the database has an impact on the domain delineation (Rafols, Ciarli & Chavarro, 2015). In the present case, it is expected that other relevant publications are to be found in more comprehensive databases, particularly regarding developing countries, as CABI or in databases with a better coverage of open access publications.

The second part of this paper intended to reveal a structure of the corpus and to compare various actors. The question of the efficiency of topic models for such a task must be raised. The complexity of topic models is balanced by its flexibility and the ability to unveil two different relationships between topics and documents. However, the variability of the output of a fitting algorithm is problematic as it has an impact on document assignment to themes and therefore on quantitative indicators of actors. Thus, quantifying this (in)stability is desirable (Hecking & Leydesdorff, 2018). As a partial answer to this issue, we fitted 10 times a 20-topic model with different seeds and calculated average country profiles to possibly challenge the results. For these 10 runs, the results for 4 themes were confirmed but, for the two themes Healthy food (Th4) and Young people diet (Th6), there was too much variability to confirm our findings. Therefore more is to be done to stabilize the themes. A solution could be to use a larger number of topics to recover the same 6 themes with refined frontiers and hopefully less variability. Such improvement is still to be experienced and its success would help to validate the findings about country profiles and interdisciplinarity indicators.

Conclusion

Despite these shortcomings, this work suggests that using topic modelling is a flexible and efficient method to answer the issue of delineating a research domain defined by large policy objectives associated to a complex societal challenge. It allows to reveal unexpected scientific approaches as well as multi- and interdisciplinary research work.

For the second purpose of revealing the structure of the domain and deriving indicators to compare various actors, topic models provide interesting analyses but with a serious cost. They allow to cross information on vocabulary similarity and on topic combination in documents. But more work is necessary to ensure the stability of the indicators derived from topics. Such a control is time consuming and, for the moment, no general method is available to achieve this control. This work therefore suggests that topic modelling is an interesting but costly method to provide a quantitative analysis of a domain.

Acknowledgements

This work is inspired by an on-going project of the OST department of Hcéres for the French Ministry of Higher Education and Research with contributions of experts in the field. Their opinions were useful to consolidate the delineation method. However, in this paper, designed as a methodological contribution, the scope has been freely restricted by the authors. Consequently, the views expressed in this publication, as well as the information included in it, do not reflect the opinion or position of the Hcéres or of the Ministry and in no way commit either of them.

References

Blei, D. M., Nag, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. Journal of machine learning research 3, 993-1022

Blei, D., M. (2012). Probabilistic topic models. Communications of the ACM 55 (4) 77-83. doi:10.1145/ 2133806.2133826

Cassi, L., Lahatte, A., Rafols, I., Sautier, P. & de Turckheim, É. (2017) Improving fitness: Mapping research priorities against societal needs on obesity. Journal of Informetrics 11 (4) 1095--1113 https://doi.org/10.1016/j.joi.2017.09.010

Di Maggio, P., Nag, M., & Blei, D. (2013) Exploiting affinities between topic modeling and the sociological perspective on culture: Application to newspaper coverage of the U.S. governments arts funding. *Poetics, 41,* 570-606. doi:10.1016/j.poetic.2013.08.004

Hecking, T. & Leydesdorff, L. (2018) Topic Modelling of Empirical Text Corpora:Validity, Reliability, and Reproducibility in Comparison to Semantic Maps http://export.arxiv.org/pdf/1806.01045

HLPE. 2017. Nutrition and food systems. A report by the High Level Panel of Experts on Food Security and Nutrition of the Committee on World Food Security, Rome. http://www.fao.org/3/a-i7846e.pdf # 12

Klavans, R., & Boyack, K. W. (2014). Mapping altruism. *Journal of Informetrics*, 8(2), 431-447. doi/10.1016/j.joi.2014.02.002

Ledford, H. (2015). How to solve the world's biggest problems. *Nature* 525, 308-311

MESRI, Ministère de l'Enseignement Supérieur, de la Recherche et de l'Innovation. (2014). SNR, Bilan de l'atelier 5 : Sécurité alimentaire et défi démographique http://www.enseignementsup-recherche.gouv.fr/cid78802/strategie-nationale-recherche-bilan-des-travaux-des-ateliers.html

Milanez, D.H., Noyons, E. & de Faria, L.I.L. (2016). A delineating procedure to retrieve relevant publication data in research areas: the case of nanocellulose. Scientometrics 107: 627. doi:10.1007/s11192-016-1922-5

Molas-Gallart, J., Rafols, I. & Tang, P. (2014) On the relationship between interdisciplinarity and impact: different modalities of interdisciplinarity lead to different types of impact. *Journal of Science Policy and Research Management* 29 (2), 69-89.https://arxiv.org/pdf/1412.6684.pdf

OECD (2017).Obesity update http://www.oecd.org/els/health-systems/Obesity-Update-2017.pdf

Porter, A.L., Cohen, A.S., Roessner, D.J., & Perreault, M. Measuring researcher interdisciplinarity. Scientometrics. 2007;72(1):117–147.

Rafols, I., Ciarli, T., & Chavarro, D. (2015) *Under-reporting research relevant to local needs in the global south. Database biases in the representation of knowledge on rice.* Available at http://digital.csic.es/bitstream/10261/132530/1/knowledgerice.pdf

Raimbault, B. Cointet,J-P & Joly, P-B (2016) Mapping the Emergence of Synthetic Biology. PLoS ONE 11(9): e0161522. https://doi.org/10.1371/journal.pone.0161522

Sievert, C., & Shirley., K., E. (2014). LDAvis: A method for visualizing and interpreting topics. *Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces* 63-70. Available at https://CRAN.R-project.org/package=LDAvis

Wallace, M. L., & Rafols, I. (2015). Research portfolios in science policy: moving from financial returns to societal benefits. *Minerva* 53 (2), 89-115. doi:10.1007/s11024-015-9271-8

Using ontologies to map between research and policy data: opportunities and challenges

Diana Maynard[1], Benedetto Lepori[2] and Philippe Laredo[3]

[1] *d.maynard@sheffield.ac.uk*
Department of Computer Science, University of Sheffield, 211 Portobello, Sheffield, UK

[2] *benedetto.lepori@usi.ch*
Faculty of Communication Sciences,Università della Svisera italiana, 6904 Lugano, Switzerland and Laboratoire Interdisciplinaire Sciences, Innovations et Sociétés (LISIS), University of Paris Est, 77454 Marne-la- Vallée Cedex 02, France

[3]*philippe.laredo@enpc.fr*
University of Paris Est, 77454 Marne-la- Vallée Cedex 02, France

Abstract

Understanding knowledge co-creation in key emerging areas of European research is a critical issue for policy makers in order to analyse impact and make strategic decisions. However, current methods for characterising and visualising the field have limitations concerning the changing nature of research, the differences in language and topic structure between policies and scientific topics, and the coverage of a broad range of scientific and political issues that have different characteristics. In this work, we discuss the novel use of ontologies and semantic technologies as a way to bridge the linguistic and conceptual gap between policy questions and data sources. Our experience suggests that a proper interlinking between intellectual tasks and the use of advanced techniques for language processing is key for the success of this endeavour.

Introduction

Mapping diverse kinds of scientific output to key topics in the science policy debate is a central concern that still requires more research. Traditional methods to characterise and visualise the field of knowledge production have important limitations. The move towards open linked data in STI studies is generating new opportunities, but also new challenges that are at the core of this paper. The opportunities concern the ability to interlink different kinds of data sources, such as publications, projects and patents, in order to provide a richer view of knowledge production (Light *et al*, 2014); the challenges are related to the need for a robust approach to identify and model relevant topics, such as those associated with specific policy and scholarly questions (Cassi *et al*, 2017).

Traditional classification systems for characterising research, e.g. the Web of Science Journal classification (Leydesdorff et al., 2009) and the IPC codes for patent classification (Debackere and Luwel, 2005), are typically simple, stable, and have widespread coverage. However, combining such schemes in order to depict an overall view of scientific knowledge production that encompasses different data sources is inherently challenging: each scheme is closely related to a specific type of data source, and despite wide-ranging efforts to map different classification schemes (Schmoch *et al*, 2003), they remain largely incommensurable. Without this overall view of knowledge production, cross-field comparisons cannot be made. Furthermore, mapping these classifications of scientific basis to policy-oriented topics presents a further issue due to terminological and conceptual divergence.

We address these problems through the use of ontologies to drive the development of a web-based tool providing interactive visualizations on European research. The tool is designed to provide information to users wishing to understand the nature of, and connections between, key European research, focusing on two topic areas central to policy makers: Key Enabling Technologies (KET) and Societal Grand Challenges (SGC). Ontologies share with

classifications the fact that they are constructed upon some intellectual understanding of reality; while their creation can be assisted by all kinds of text-based methods, they ultimately require some method of expert-based arbitration and must rely on some kind of "shared vision of the structure of the domain of interest" (Daraio *et al*, 2016). Our experience shows that while natural language processing (NLP) techniques are critical for linking ontologies with large datasets and extracting from the latter robust evidence, nevertheless some key design choices on the ontology and its application to data are basically of an intellectual nature. This suggests that the design of robust interactions between expert-based priori knowledge and evaluation on the one hand, and the use of advanced data techniques on the other hand, is a key requirement for robust S&T ontologies.

Related Work

A large body of work has been developed to address the limitations of classification systems outlined above. These mostly rely on individual data items, and include citation analysis (Šubelj *et al*, 2016) and NLP (Van den Besselaar and Heimeriks, 2006). Recent NLP work has focused on extracting relevant information from scholarly documents[1], but this is primarily concerned with metadata and citation extraction. Other research has investigated keyword extraction from academic publications (Shah et al., 2003) and overlay maps (Rafols, 2010). The semantic web approach of Motta and Osborne (2012) in Rexplore takes scholarly data analysis a step further by examining research trends at different levels of granularity, and by finding semantic relations between authors, using relations such as co-citation, co-publication and topic similarity. However, this is again limited to only publication data, which is relatively cohesive.

Another strand of research that moves away from traditional classification systems involves modelling topics and domains in order to gain an overview of the field. Here, techniques such as LDA (Blei et al., 2003), PLSA (Blei, 2012) and KDV (Börner *et al*, 2003) are used extensively for understanding and mapping large research areas, for example to understand the evolution of topics over time (Chen et al., 2017). These techniques essentially model the distribution of topics, based on the principle that documents typically contain multiple topics according to a probabilistic distribution. However, the drawback is that it can be hard to make sense of the resulting information and to understand the nature of these unlabelled clusters and topics, and this work often has to be done manually. Furthermore, these methods do not deal well with topics outside a core subject domain, and clustering may result in multi-disciplinary topics without strong internal cohesion (Boyack, 2017).

All these techniques extract topics in a bottom-up manner from structural (in the case of citation analysis) and linguistic (in the case of NLP and topic modelling) features of documents. However, while these provide detailed views of specific knowledge domains and of their evolution over time, they are currently less suited for large-scale mapping across the whole S&T landscape. Methods like LDA also work well on homogenous kinds of data, such as a large collection of publication abstracts, but cannot extract good topics that encompass the diversity of more disparate datasets. Finally, connecting such topics with relevant themes at the policy level is far from simple, since the associated terminologies are largely incompatible (Cassi *et al*, 2017).

Task

In this work, we use an ontology-based approach to address these issues. An ontology is defined as the "explicit formal specification of the terms in the domain and relations among them" (Gruber 1993), and in our case, acts as a bridge between the different and evolving vocabularies across heterogenous data sources. Our ontology is essentially a hierarchical

[1] http://csxstatic.ist.psu.edu/about/scholarly-information-extraction

representation of topics, as seen in many traditional classification systems, but with the possibility of multiple inheritance. While keeping some basic features of classifications, like the presence of a core set of subjects organized in layers, ontologies are more flexible in their structure and, through instances (keywords), can be connected to (different and evolving) vocabularies across heterogeneous data sources. An ontology thus offers a formal representation of a domain of knowledge that is shared by a specific group of "experts", based largely on a priori conceptual knowledge, while clustering approaches represent a more informal, data-driven view.

Challenges

Implementing an ontology involves 2 major aspects: first, the design of the ontology structure, consisting of a set of related topics and subtopics in the areas of KET and SGC; and second, a way to map relevant documents to topics in this structure (which can be seen as a problem of multi-class classification, with a large number of classes).

There are several major challenges associated with the design and use of ontologies in this scenario: both conceptual and practical. First, the ontology structure is hard to define because it is not clear what level of precision is both needed and practical, and because these affect the implementation of the document-topic mapping. The structure must also be intuitive for human users to navigate, and this is perhaps the most challenging component: it must reflect both the needs of the policy makers, but also the variety of ways in which information is portrayed in the data sources (in our case comprising publications, project abstracts, patents, and descriptions of social innovation projects). This latter is also critical for the data classification task.

We have attempted to mitigate this problem by consulting experts at every stage of the process, holding workshops with policy makers from a variety of fields in order to understand their needs, and following a principled development process. However, the intrinsic vagueness of the notion of Key Enabling Technologies and especially Societal Grand Challenges means that the topics are hard to define, and there is no gold standard against which to evaluate.

Second, differences in vocabularies within academia, industry and society mean that the same concepts are typically expressed in very different ways, especially in patents which are extremely technical. Existing attempts at classification, as described earlier, have highlighted these issues. Our solution to this lies in the use of sophisticated techniques from NLP and Machine Learning, where this kind of language variation is a common problem and techniques go far beyond the simple keyword matching approach used in other work. For example, word embeddings and a plethora of similarity and distance measures are used to determine possible mappings between data and classes in the ontology.

Finally, there are numerous issues related to evaluation of such a large-scale classification. It is impossible to know if every document has been correctly classified, and almost certainly there will be errors. We mitigate this by testing in different ways and at a variety of stages in the process, checking a sample of annotated documents, looking at the global picture for incongruities (such as topics with an unexpectedly high or low number of documents), evaluating different keyword generation strategies, and tweaking the ontology where needed.

Approach

The method we adopt in this work comprises 3 steps: ontology creation, ontology population, and ontology-based classification (data annotation). All three steps require human intervention to define prior assumptions and to evaluate outcomes, but they integrate automatic processing through advanced language analysis techniques. Consequently, if any changes are deemed necessary, the process can easily be rerun and the data re-annotated within a short period of

time and in a principled way. The current version of the ontology contains 150 topics based around the 6 KETs and 7 SGCs, and around 8,700 unique keywords.

Ontology creation

The ontology is defined according to the two strands of KET and SGC. We take as a starting point some existing classifications, which we merge and map, such as the mappings between IPC (International Patent Classification) codes and both KETs (Van der Velde, 2012) and SGCs (Frietsch et al., 2016). For KETs, we also make use of the structure implemented in the nature.com ontologies portal (Hammond and Pasin, 2015). Some of these topics are already connected to DBpedia and MESH, which provides us with an additional source of information for keywords. We manually refine this structure, removing the lower levels, to make a slightly more generic set of topics. We also create subclasses based on EU policy documents, which describe how the KETs and SGCs are structured. A key expert decision relates to the extent of overlap between classes and subclasses, as some KETs are intrinsically related.

Ontology population

Having created an initial structure containing the concepts (topics and sub-topics), the ontology then needs to be populated with instances (keywords) from various data sources. These instances help us to: (1) match user queries to topics in the ontology; and (2) match documents from the various databases to these topics. These two issues form the crux of the system.

The first stage consists of automatically generating key terms from the ontology class names and associated information, such as class descriptions, using Automatic Term Recognition techniques. Additional terms are manually generated by experts where information is sparse or where there is possible ambiguity. Terms in which the experts are highly confident are designated "preferred" and are used as seed terms for the expansion stage. These are typically the topic name itself, synonyms or linguistic variants of it, and additional manually generated terms. For example, one preferred term for the topic "intelligent transport" is "intelligent navigation". The remaining (non-preferred) terms are the automatically generated ones, and are only used for the matching stage later. These have a lower weighting during the matching, since we are less confident about their relevance. An example of a non-preferred term for the topic "intelligent transport" is "radar tracker" (which is somehow connected with the topic but is not a close synonym). This term might be relevant if found in conjunction with another relevant term for the topic, but not on its own.

The second stage involves the generation of additional keywords. First, our preferred terms are used to generate a seed set of initial keywords associated with each ontology class. We then find semantically similar terms to these using word embeddings trained on a large corpus of just over 8.3 million documents, comprising a mixture of publications, project descriptions, policy documents and patent abstracts. This corpus will be extended periodically as additional data becomes available – while larger corpora may provide better training, there is a tradeoff between this and the relevance of the documents (our previous experiments showed that using larger corpora of pre-trained embeddings on more general corpora gave worse results). The method consists of extracting a set of possible terms from that corpus using Automatic Term Recognition and NLP techniques, and then finding the ones most similar to the seeds. Finally, the terms are scored according to how "representative" they are of that class, and prior probabilities are generated using PMI for term combinations, based on frequency of co-occurrence in the training data.

The implementation of this process showed that automatic techniques enable the generation of a large number of keywords, but become problematic when two subclasses share some similar terms (like rail and road transport). Additional statistical techniques can be used to further weight terms based on maximising the semantic distance between terms from such closely

related classes, but some level of expert intervention is nevertheless required in order to delimit the subclasses and to attribute a sufficient number of distinct terms to each of them.
The result of the ontology population stage is thus a set of keywords associated with each class, each of which has a score indicating the degree of its relevance to that class. These keywords are used for the mapping between documents and topics in the final data annotation stage.

Data annotation

The data sources take the form of 4 databases containing information about projects, patents, publications and social innovation respectively. The idea of the annotation is to link each data element (e.g. a project) with the relevant topic(s) in the ontology, so that indicators (and from there visualisations) can be built around these. For example, in order to know how many EU projects there have been about "gene therapy" in a particular year and location, we must first know which projects should be associated with this topic.
We have developed a classifier which takes documents as input and returns information about the class(es) to which each is linked, and a score for it. The scores are based on (i) the weight of that keyword for that class (e.g. preferred terms have a higher score, as do terms ranked close in similarity to these); (ii) the combination of keywords found in the document using PMI calculations from the ontology population stage (on the assumption that term combinations with high PMI are better indicators that the class is relevant for that document); (iii) subclass boosting, whereby keywords belonging to a more specific class in the ontology are to be preferred over more general ones.
The process of classification thus assigns multiple possible topics to each document, not all of which are likely to be useful as some will be low-scoring. Thresholds are used to decide which of the topics are most relevant, based on analysis of distributions and some inspection of results. This is a typical expert-based task that involves manual checking of classified documents to find a reasonable balance between recall and precision.

Discussion and Conclusions

In this work, we aim to address some of the current limitations in applying traditional classifications to a science policy domain, through the use of ontologies, thereby extending the reach of existing text-based methods while still maintaining the power and rigour of classification systems. In particular, we overcome the problems in connecting policy-based topics with science-based topics. This wide-ranging view of the research domain requires the focus to shift from static maps and detailed analyses towards indicators that can be compared temporally, geographically, and across topics.
Our approach is designed to maximize automated processes wherever possible, which is not only critical for dealing with massive volumes of data, but also lends itself to domain and topic adaptation. Since research is not static – topics change over time, new terminology comes to the fore, and even geographical boundaries do not remain the same – this enables much greater flexibility than many existing classification-based systems. Changes to the ontology or the input of new research data can easily be handled automatically, and updates pushed seamlessly to the central databases from which visualisations are generated. On the other hand, these are tempered by expert intervention at critical stages.
There are, however, limitations. Rigorous evaluation is always difficult, and requires manual intervention, which is time-consuming and subjective. The use of NLP techniques brings its own problems: language is tricky for machines to understand, and tools will never be 100% accurate. Numerous issues in terminology extraction still need to be solved globally: many terms are ambiguous and require at the least context, and in some cases, only the kinds of world knowledge that humans can provide. Nevertheless, this work provides some critical new pathways for STI technologies, which open up avenues for future research directions.

Acknowledgements

This work was partially supported by the European Union under grant agreement No.726992 KNOWMAK and grant agreement No. 825091 RISIS.

References

Blei, D.M., Ng, A.Y. and Jordan, M.I., 2003. Latent dirichlet allocation. *Journal of machine Learning research*, *3*(Jan), pp.993-1022.

Blei, D.M., 2012. Probabilistic topic models. *Communications of the ACM*, *55*(4), pp.77-84.

Börner, K., Chen, C. and Boyack, K.W., 2003. Visualizing knowledge domains. *Annual review of information science and technology*, *37*(1), pp.179-255.

Boyack K (2017) Investigating the Effect of Global Data on Topic Detection. In l¨aser, J., Scharnhorst, A. & Gl¨anzel, W. (eds), Same data – different results? Towards a comparative approach to the identification of thematic structures in science, *Scientometrics*, *111*(2), 2017, pp.999-1015.

Cassi, L., Lahatte, A., Rafols, I., Sautier, P., & De Turckheim, E. (2017). Improving fitness: Mapping research priorities against societal needs on obesity. *Journal of Informetrics*, *11*(4), 1095-1113.

Chen, C. (2017). Expert review. Science mapping: a systematic review of the literature. *Journal of Data and Information Science*, 2(2), 1-40

Daraio, C., Lenzerini, M., Leporelli, C., Moed, H. F., Naggar, P., Bonaccorsi, A., & Bartolucci, A. (2016). Data integration for research and innovation policy: an Ontology-Based Data Management approach. *Scientometrics*, *106*(2), 857-871.

Debackere, K., & Luwel, M. (2004). Patent data for monitoring S&T portfolios. In *Handbook of Quantitative Science and Technology Research*(pp. 569-585). Springer, Dordrecht.

Frietsch, R., Neuhausler, P., Rothengatter, O., Jonkers, K.: Societal grand challenges from a technological perspective: Methods and identification of classes of the international patent classification IPC. Tech. report. Fraunhofer ISI Discussion Papers Innovation Systems and Policy Analysis (2016).

Gruber, T. (1993). What is anOntology. *http://www-ksl. stanford. edu/kst/whatis-an-ontology. Html*

Hammond, Tony, and Michele Pasin. "The nature. com ontologies portal." *5th Workshop on Linked Science,* 2015.

Leydesdorff, L., & Rafols, I. (2009). A global map of science based on the ISI subject categories. *Journal of the American Society for Information Science and Technology*, *60*(2), 348-362.

Light, R. P., Polley, D. E., & Börner, K. (2014). Open data and open code for big science of science studies. *Scientometrics*, *101*(2), 1535-1551.

Maynard, D. and Lepori, B. Ontologies as bridges between data sources and user queries: the KNOWMAK project experience. *STI 2017*, Paris, France, September 2017.

Motta, E. and Osborne, F. Making sense of research with Rexplore. In *Proceedings of the 2012th International Conference on Posters & Demonstrations Track*-Volume 914 2012 Nov 11 (pp. 49-52). CEUR-WS. org.

Rafols, I., Porter, A.L., Leydesdorff, L.: Science overlay maps: A new tool for research policy and library management. *Journal of the American Society for in- formation Science and Technology* 61(9), 1871–1887 (2010).

Schmoch, U., Laville, F., Patel, P., & Frietsch, R. (2003). Linking technology areas to industrial sectors. *Final Report to the European Commission, DG Research*, *1*(0), 100.

Shah, P. K., Perez-Iratxeta, C., Bork, P., & Andrade, M. A. (2003). Information extraction from full text scientific articles: where are the keywords?. *BMC bioinformatics*, *4*(1), 20.

Šubelj, L., van Eck, N. J., & Waltman, L. (2016). Clustering scientific publications based on citation relations: A systematic comparison of different methods. *PloS one*, *11*(4), e0154404.

Van den Besselaar, P., & Heimeriks, G. (2006). Mapping research topics using word-reference co-occurrences: A method and an exploratory case study. *Scientometrics*, *68*(3), 377-393.

Van de Velde, E.: Feasibility study for an EU monitoring mechanism on key en- abling technologies. IDEA Consult (2012.

Innovation policy and governance networks on national innovation systems

Luis Antonio Orozco[1], José Luis Villaveces†, Gonzalo Ordonez-Matamoros[2] and Gabriel Moreno[3]

[1] *luis.orozco@uexternado.edu.co*
School of Management – Universidad Externado de Colombia - Bogotá (Colombia)

[2] *gonzalo.ordonez@uexternado.edu.co*
FIGRI – Universidad Externado de Colombia - Bogotá (Colombia)

[3]*morenoluis@javeriana.edu.co*
School of Computing Engineering - Pontificia Universidad Javeriana - Bogotá (Colombia)

Abstract
National innovation system - NIS became as a social construction that can be represented using social network analysis. Several countries created national agencies to administrate resources for research in order to achieve competitiveness and recently, sustainability. Research on NIS stress the importance of coordination between agencies and actors. However, policy networks as a channel to create spaces from cross-agency collaboration toward democratic participation in governance networks is one of the topics neglected in the literature, taking into account the role of the agencies devoted to managing innovation policies. This study used social network analysis to evaluate centrality measures to show the behaviour of several agencies in the definition, execution and discussion of innovation policy in Colombia, defining bargaining subsystems in terms of law, by investment of resources and by public deliberation and accountability in online social networks like Twitter, in order to contribute to a theoretical and measurement proposal of how policy networks evolve to governance in NIS.

Keywords: Policy Network, Governance Network, Innovation System, Colombia

Introduction

Policy Networks -PN in general and cross-agency collaborations in particular respond to increasing complexity of problems public administrations face, representing important governance challenges described in theoretical (Hudson et al., 1999) and in empirical studies (Yun, Ku and Han, 2014). In the development of PN, public entities harness complementarity and achieve better results thanks to synergies created in the coordination of policies at program levels. Often, however, agencies compete amongst themselves to accomplish their organizational mission and show results to gain recognition and legitimacy, increasing their budget (Hudson et al., 1999), reaching points where boundaries of organizational roles become blurred when complex affairs need to be tackled, and overarching goals fulfilled (Daley, 2009). In this process, tensions, lack of alignment and even conflicts between agencies emerge. This is particularly true when institutional roles are developed to design and implement national policies in topics that cut across various sectors and aim to address issues as complex as innovation and entrepreneurship where, as Audretsch claims, "aspects relevant to entrepreneurship policy can be found across a broad spectrum of ministries and agencies, ranging from education to trade and immigration" (Audretsch, 2004, p. 183). Innovation policies networks, conceived as part of national innovation systems – NIS, imply the creation

of several types of interactions among different types of actors, where cross-agency collaboration is central to develop governance networks – GN, which in turn imply the inclusion of interests and stakeholder groups active in a democratic system, open to public scrutiny and participation.

This paper builds on governance studies aiming at better understanding the challenges and operationalization of the NIS concept, striving to overcome the limitations of narrow normative claims by offering alternative perspectives related with network development and policy and programmed coordination (Kuhlmann and Shapira, 2006, Laranja, 2012; Chaminade and Padilla, 2017, Dutrénit and Puchet, 2017, Lindner et al., 2016). In so doing and acknowledging the need for more theoretical and conceptual models and elaboration in this relatively new topic, this paper offers a novel approach and an empirical analysis of a particular national case to add to current understanding of the ways both innovation PN and GN concepts work in general, and in an emerging economy context (Karo, 2010, Kuhlmann and Ordonez-Matamoros, 2017) in particular: Colombia.

The research questions this paper attempts to address are therefore: what are the main features characterizing Colombian NIS and its evolution from the perspective of the PN and GN concepts? What is the specific position Colciencias (the Colombian National Science Foundation) holds in such networks? We performed a multilevel system analysis, and studied a) the legal mandate promoting the creation and operation of interagency coordination involving high-level decision-making councils, b) programmatic actions implemented by agencies coordinating their efforts to support innovation, and c) online deliberate interactions of agencies and actors engaged in public discussions on innovation policies as reflected in the Twitter social network.

After analyzing data on the three levels mentioned from the period 2012 to 2014, we found that Colciencias holds a central position in the PN of the Colombian NIS, despite intense competition with several agencies that have budget and legal positions to develop innovation policy and programs, the legal disorder for innovation policy and debates over its legitimacy, the control and the organizational capacity of Colciencias as a main rector of NIS.

This paper is structured as follows: in the next section, the literature review on policy and governance networks in particular and on NIS in general is presented. The third section presents the methodology implemented to answer the aforementioned research questions. The results of the Colombian' case is discussed in the fourth section. The last section offers the conclusions and discussion while highlighting and proposing a future research agenda.

Literature review

Innovation policy represents an important challenge because it has not only to address the science and technology advancement itself, but also other domains like education, agriculture, labor, industry, health, transportation, information and communications technology, among others (Laranja 2012, Kuhlmann and Shapira, 2006). Innovation policy has become a horizontal policy that could complement, and in other cases affect, the traditional policy areas, involving conflicts between diverse interests, external pressures and threats of traditional hierarchical organization of the state (Karo, 2012), as showed in biotechnology in Colombia (Orozco et al., 2007).

The model of NIS in Northern countries emerged in the recognition of the dynamics developed in the interaction between public and private entities that created network´s structures that steer innovation. It was a bottom up phenomenon. In Latin American countries, in contrast, the introduction of the NIS notion became a top down policy trend that tried to replicate the dynamics in the North (Alcorta and Perez, 1998, Arocena and Sutz, 2002). For example, in the Colombian case the promotion of innovation policies lacked the awareness and participation of industry and academy, despite the efforts of public agencies like Colciencias, as a rector of S&T

policy (Villaveces, 2002). It has been claimed that policies objectives in the frame of Latin American NIS are beyond the administrative capacity of a single public agency, despite their legal competence to do it (Alcorta and Perez, 1998, Casas, Corona and Rivera, 2014).

Also, NIS lacks a "proper theory for the role of the state" (Karo, 2012, p. 496). Innovation system studies emphasize interactions between different stakeholders without specific attention to the internal logic and processes of policy-making (Karo, 2012) and the development of PN, while GN lacks a comprehensive theory to explain under which conditions it may offer advantages (Laranja, 2012). Beyond these issues, is not clear how the main public agency created to steer innovation policy, collaborate and compete in the development of PN and GN.

PN in NIS traditionally determine the political agenda in the hierarchical line when higher-level councils define guidelines and public agencies in the middle level execute them. This hierarchical reasoning involves a tension. Nations created national agencies to steer and fund innovation, and managerial autonomy is needed for experimentation and adaptive policies to support private sector risk-taking and experimentation in innovation policies (Karo, 2012, Lundvall et al., 2009). This is difficult to achieve when there are PN shortcomings as discussed above, and different measures of performances for different public agencies ranged for patenting and publications from socio-economic indicators. Also, the participation of different stakeholders in several ways represents challenges for public agencies that reach to develop GN.

The term governance was introduced into innovation policy studies recently in moving beyond the linear model of science and technology policies. It includes systematic thinking in the process of complex negotiation between different systems like science and technology, industry, civilians and non-governmental organizations as well as public administration in the prioritization of topics and the construction of a shared vision of the future (Kuhlmann and Shapira, 2006). Also, the control in the allocation of tasks and budgets implies for governance the accountability of PN and the development of different channels of information to enhance public monitoring and participation on innovation policies (Laranja, 2012).

The main objective of NIS is to promote networks and achieve synergies from their joint operation. PN could be able to regulate interactions, provide incentives and reduce uncertainty (Lundvall et al., 2009). Then, cross-agency collaborations as part of PN are devoted to harmonizing different competences and contributions of each agency to provide an institutional space that facilitate the relationships and the development of innovations in the inclusion of different stakeholders allowing mechanisms like ICT social networks to engage in the construction of GN.

In the frame of GN, experts like academicians, labor union leaders, consumer groups, civilians, and entrepreneurs participate in the councils and intermediate PN in their decision-making and execution of innovation policies. Using semi-formal coordination based on dialogue, negotiation and collective deliberations with ministries and stakeholders, GN promote discussions and accountability in the setting and achievement of common goals in NIS. Therefore, affairs like innovation and entrepreneurship demand the creation of PN able to incorporate civil society in the development of GN to improve the process of translating policies into improvements in economic and social affairs.

PN in the governance model steer the division of tasks between different stakeholders and, in so doing, create roles that are more blurred and flexible. "From the perspective of innovation policy discourse, the more networked and stakeholder inclusive mode of policy-making is supposed to increase the speed, flexibility, stakeholder commitment and the efficiency and adaptability of policy-making" (Karo, 2012, p. 505). However, steering PN to create GN can reduce the actual policy power to create consensus and common points of view. Stakeholders embedded in policy networks squeeze bureaucratic competences of public agencies and could

paralyze the development of policies (Karo, 2012). The behaviour and contributions of many agencies and several actors in a big size network are difficult to manage.

The GN do not replace hierarchical organizations and PN, but rather co-evolve and complement, adding value to the efforts of coordination and deliberation in the public realm. Accountability and enforcement not only use traditional mechanisms of command and control, but also create self-controls by relying on peer evaluation to reinforce commitment to perform common goals on innovation policies that depends on visibility and perceptions in the public domain (Laranja, 2012). As discussed above, the mass media and public communications in ICT channels like Twitter and Facebook enhance the participation and awareness of different stakeholders. As such, the traditional and bureaucratic forms of control are enhanced by the ICT that enables the participation of the public.

Both PN and GN would be expected to reinforce each other. However, innovation policies designed by the rector agency in ST&I system and in national councils often propose goals and means, and later, must develop PN to bargain with several agencies and do lobby in the congress to introduce legislations that enables their programs. Then, the power to lead innovation PN could shift from a public agency to another due to their legitimacy and recognition in GN. Changes in laws, agencies budgets and public recognition and acceptance could change the legitimacy of national agencies created to steer innovation policy. In the next section, we propose a methodology in which we describe the evolution of the NSI in Colombia and evaluate the role of Colciencias, as main public agency in the Colombian NIS, that must be able to be central in PN that evolve in GN using online social networks to steer policies as a main actor that is able to connect several agencies and actors in three levels of governance.

Methodology

We attempt to address two research questions: a) what are the main features characterizing Colombian NIS from the perspective of the PN and GN?, and b) What is the specific position Colciencias holds in them? To answer to this question, we performed a multilevel system analysis, and studied a) the legal mandate promoting the creation and operation of interagency coordination involving high-level decision-making councils, b) programmatic actions implemented by agencies coordinating their efforts to support innovation, and c) online deliberate interactions agencies and actors engage in public discussions on innovation policies as reflected in the Twitter social network. To study the NIS in Colombia with the aim to identify cross-agency collaboration and the role and structural position of Colciencias within the system, we combine three theoretical frameworks to propose three subsystems in which we can study the rationale of PNs evolving towards GNs.

The Viable System Model developed by the organizational cybernetics field has been applied to understand the NIS in Colombia (Sánchez and Pérez, 2013) and can be used to separate process. According to Beer (1984) the viable system model encompasses three levels of actions. The strategic system defines the policy to organize, structure and control the actions toward the goals defined. The execution subsystem includes the activities performed by actors and their coordination. Finally, the control subsystem evaluates the execution according to the policy, given feedback of the performance to adjust policies and execution subsystems.

In addition, in the development of PN as a process to define, steer and evaluate public policies, Fountain (2002) defined three stages in the process of cross-agency collaboration: the production process that consist in bargain and negotiation of goals and responsibilities, the integration process that defines the execution of policies with shared resources and goals, and the accountability, that relies in the evaluation of joint actions between agencies and other actors that are able to evaluate and discuss the results.

Finally, the Theory of Organizational Institutionalism (Scott, 2005) help us to explain the process in which PNs seek to gain consensus and enact procedures to manage conflicts while

creating highly resilient structures in the long term. Then, PNs act to achieve desirable results that are consistent with laws, norms and the cultural cognitive realm of society. The law, as a first pillar of institutions, enforce the behaviour of organizations stating the mandatory roles and actions. Norms, as a second pillar, defines behaviours that are considered right in the moral realm of an organizational field, and several agencies certified the compliance of an organization that could be considered as a member of an elite (for example, becoming an OCDE member). The cultural-cognitive pillar defines the behaviour and actions that are taken for granted in a society, and represent the public opinion over organizational practices, and the general acceptation or rejection of an organization in an institutional realm (Scott, 2005).

Then, the evolution of PNs to GNs can be understood and researched by the viable system model proposed by Beer (1984), using the rationale of cross-agency collaboration defined by Fountain (2002) in the framework of three pillars of institutionalization (Scott, 1995). We defined three subsystems:

The first one is the regulatory system defined by laws that organize councils composed by different public and private agencies that must meet to define national policies. This encompasses the strategic system defined by Beer (1984) and the production function defined by Fountain (2002) in the regulatory pillar of institutions (Scott, 2005). The regulatory frame defines formal memberships of PN with specific roles and actions that may be traceable and accountable in a governance process (Fawcett and Daugbjerg, 2012). High level councils may act as a coordination body in which system' goals are defined and traced. Memberships of PN may be defined by law in terms of councils and other kinds of boards that determined the first instance of inter-agency collaboration that sets the strategy in a process of bargaining and negotiation.

In this framework, the PN is constructed by the relationship between agencies that compose the boards according to three regulations aiming at developing innovation policy in Colombia: Law 1286 of 2009, which created an advisory council with several agencies; Law 1530 of 2012, which created the so-called OCAD board, composed by national and regional agencies with the participation of universities, where decisions upon STI projects to be funded by the mining royalties in Colombia are made; and Decree 1500 of 2012 which defines a national council for competitiveness and innovation, composed by 45 public and private actors, including public agencies, universities and firms that define policies and programs to steer and promote industrial development in the country. An agency or an actor defined by regulations that belongs to the same board with others present a relationship creating a PN.

The second one is a programmatic system. This instance encompasses the execution subsystem (Beer, 1984) in the light of normative pillar of institutions in which organizations align their actions to become as an elite member that behave in a morale realm (Scott, 2005) in the development of an integration process (Fountain, 2002). This process refers to the mechanisms in which different agencies embedded in PN bargain or negotiate the definition of programs, common and independent responsibilities, competencies, resources and/or other organizations that are needed to enact policies, as well as ways of ensuring accountability. Recognizing that public agencies could compete with each other to achieve specific goals, gain recognition and justify budget, the bargaining process that result in common activities and programs could be traced by joint calls and activities. These actions represent the interchange and complementarity achieved in such PN. This collaboration might be based on personal relationships between public managers and/or on signed agreements between agencies. Membership of PN may be defined by agencies that alone or with collaboration design and execute programs to fund and promote innovation. To construct this network, we looked at the Colciencias database on projects and programs that ran between 2010 and 2014, relating to agencies that fund innovation activities. In this context, a relationship is defined by the participation in mechanisms used to promote innovation in Colombia.

The third instance, called Online Social Network System - OSN, encompasses the coordination and control mechanisms (Beer, 1984) in the realm of cultural-cognitive pillar defined by Scott (2005) in which civil society participated in the definition, execution and accountability of PN performance (Fountain, 2002). When PN expand their actions and influence other stakeholders, they begin to stablish relationships in the public discussion of policies. Factors that can determine the way that power is organized and governance is carried out are society's capacity to be informed through mass media as well as it's tendencies to obey, reject, or be indifferent to policy incentives, actions and consequences. The public opinion evolved in mass media and social networks allow the construction of a perceptual imaginary that construct the GN. Mass media and public opinion are the key channels to construct and define legitimacy (Deephouse and Suchman, 2008). In a democracy, the change or maintenance of rules and practices needs authority that can only be built on legitimacy. As such, public opinion is the key element supporting the institutionalization of a governance structure (Schepers, 2013) and can be used in a way to measure the role of several agencies in PN in the development of GN. Using Twitter as a channel of public discussion, we can construct a GN with messages emitted between agencies defined by regulatory and programmatic systems and organizations and persons that engage and participate in their messages.

To construct this network, we identified the official Twitter account of each agency that belongs to the three boards defined in the regulatory system. Using data mining techniques filtering messages with the word innov*, we defined a relationship when one account refers another in their messages between 2010 to 2014. For example, if the account of Colciencias refers in one message that contains the word innov* other agency or actor account, we interpret that there is a relationship. This includes not only the agencies defined by regulatory system, but also accounts that, talking on innovation, refers other actors.

To measure the position of each agency and actor involved in the three subsystems, we use social network centrality measures: degree, closeness, betweenness (Wasserman and Faust, 1994). Degree measure computes all direct relations to evaluate prestige of each actor. "The simplest definition of actor centrality is that central actors must be the most active in the sense that they have the most ties to other actors in the network or graph" (Wasserman and Faust, 1994, p. 178). Closeness measures the ability to presents reciprocal links. "The measure focuses on how close an actor is, to all the other actors in the set of actors. The idea is that an actor is central if it can quickly interact with all others." (Wasserman and Faust, 1994, p. 183). Betweenness measure evaluates the capacity to connect different nodes in the networks. "An actor is central if it lies between other actors on their geodesics, implying that to have a large 'betweenness' centrality, the actor must be between many of the actors bia their geodesics" (Wasserman and Faust, 1994, p. 189), that means "a shortest path between two nodes" (Wasserman and Faust, 1994, p. 110).

Results

Regulatory system

Law 1286 of 2009 reorganized Colciencias out of National Planning Department - DNP trying to achieve a ministerial level for their dependence of the presidency, and defined it as the rector of the new science, technology and innovation system. This law assures resources from royalties in mining and oil extraction to be invested on scientific and entrepreneurial activities to promote innovation. Also, this law defines a PN between agencies that can bargain goals and resources to steer innovation. The law 1530 of 2012 stated a new national board called OCAD to evaluate and decide the funding for projects presented by the 32 regions of Colombia. This council is composed by the DNP, Colciencias and regional universities and governors. The idea of this PN is to enhance the relationships between local, regional and national-central authorities

to decide the investment of royalties in research and innovation. Finally, the decree 1500 of 2012 created a national council for competitiveness and innovation, composed by 45 public and private actors, including public agencies, universities and firms that define policies and programs to promote the industrial competitiveness.

These three regulations could imply competition in the construction of PN and governance mechanisms on innovation policies. Using the agencies involved on these regulations, we evaluated the centrality for each agency. Figure 1 presents the relationships between the actors that composed the advisory council of STI defined by Law 1286 of 2009, composed by 17 actors: 4 ministers (Education - MEN, Commerce and industry - MCIT, Agriculture - MADR, and Work - MPS); 3 public agencies (Colciencias, Sena (technical public training agency) and DNP); 6 academics and 4 entrepreneurs. The Law 1530 of 2012 composed the OCAD by 19 actors: 6 governors, 6 universities, 5 ministers (MEN, MADR, Information and communication ministry - MINTIC, Mine and Energy - MINMINAS and Treasury - MINHACIENDA) and 2 agencies (Colciencias and Sena). Finally, the Decree 1500 of 2012 composed the national council for competitiveness and innovation by 45 actors from public and private sectors.

In Table 1 we present the measures of degree, closeness and betweenness centrality for the main agencies in the network. Colciencias, MEN, MADR and DNP presents the highest values of centrality in all three measures. These agencies participate in the three councils defined by law and bear the responsibility, implicit, to coordinate the policies to steer the NIS. Contrary to the idea that Colciencias coordinate and orchestrate the PN, this analysis implies that it shares such roles with MEN, DNP and MADR in traducing the interest of several agencies into general policies that define programs and other activities discussed below in the programmatic system.

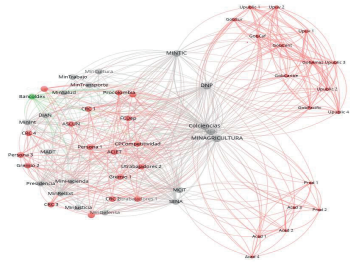


Figure 1. Policy network of Law 1286/2009; Law 1530/2012 and Decree 1500/2012.

Programmatic system

Compared with the conventional hierarchical organization, managing cross-agency arrangements requires trust in order to coordinate the strategies of participating agencies with different goals and preferences (Lips et al., 2011). In problem-solving and information sharing, agencies can develop joint programs that represent a good indicator for cross-agency collaboration. Table 1 presents the centrality of each agency in which Colciencias, Sena and MINTIC lead the cross-agency collaboration in the development of funding mechanisms to steer innovation policy in Colombia. As for the programmatic system, composed by agencies that promote innovation, and as presented in Figure 2, Colciencias appears to play a central role in channeling joint efforts and resources for innovation projects. In a second place, Sena, which main mission is technical training, also appears to be an important bridging actor in the PN. Apparently, other key actors do not use Colciencias as channel of their resources for R&D activities. In fact, the MARD develops their policies without direct relationships with Sena and

Colciencias, which may imply a disruption of systemic efforts. Interestingly, multilateral agencies as InterAmerican Developing Bank - BID and World Bank - WB are nodes that, due to their centrality, seem to coordinate the relationships between Colciencias, Sena and MARD. This result supports the later description that proposed that powerful ministries like agriculture develops their own policies on innovation.

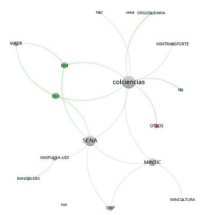


Figure 2. Policy networks in programs and projects between 2010 to 2014. Source: Colciencias Innovation Database 2010-2014.

Online Social Network system

The next result shows the relationships created by the public discussions between agencies in the Twitter social network between 2011 to 2014. In Figure 3 we can see a center-periphery network in which a few agencies lead the public discussion about innovation, while several actors show a few interactions with the main agencies. In Table 1 Colciencias is the most central agency in public discussions of innovation in all centrality measures, ahead of the President Santos account. MEN ministry, Santiago Rojas, the director of the MINCIT and Yaneth Ghia, director of Colciencias appeared as central nodes in the OSN.

Although one would expect to see the most important agencies of the regulatory and programmatic systems playing important roles in the OSN (ministries and the DNP), one in fact finds that they barely appear discussing on innovation in Colombia via Tweeter. It is in fact DIAN (national tax agency), the presidency, Bancoldex (Colombian agency to fund exports and international commerce) and Sena who, together with Colciencias, play such role.

In terms of centrality degree, Bancoldex and MCIT present the greatest number of communications. However, Bancoldex and Innpulsa (agency created to invest on entrepreneurship and innovation), that are not included in the regulatory and programmatic systems, have been gaining a public recognition as agencies that can develop policies and direct programs on innovation, according to centrality measures in Table 1. Sena is important for its capacity to interconnect communities, despite their low grade of direct ties. It is important to highlight that universities like Unicauca, industrial unions like ANDI, the private association of industrial enterprises, Federación Nacional de Cafeteros for third party association and Cofecamaras, the association of Chambers of Commerce, Governors of Chocó and Vichada, appeared as an important actors in the discussion on innovation affairs, despite they does not appeared in the top 10 agencies showed in the Table 1. This reveals the inter-organizational heterogeneity on networks that increase the capacity to perform better the NIS in terms of discuss and be aware of policies and resources.

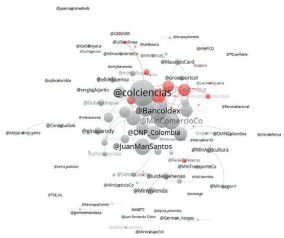


Figure 3. Online social network system. Source: Twitter Database 2010-2014.

In Table 1 we present the comparisons of the three systems discussed above. The main conclusion is that Colciencias remains as the most central agency in the Colombian' NIS. DNP also is the central agency in PN and OSN in the NIS. The MEN is important in regulatory system, but not in the programmatic and governance system. This could imply a failure in the development of PN towards GN if MEN defined the policies of higher education system, that represents near the 85% of participation in STI activities in Colombia (Villaveces and Orozco, 2015). Sena is the second main agency in terms of centrality in the programmatic system. This implies that coordination between Colciencias and Sena is the principal way to execute policies and link research with industrial needs. MICIT and Innpulsa appeared to be agencies that could complement or compete with Colciencias in the public discussion of innovation. However, their participation in programmatic system is low compared by other agencies like MINTIC. Finally, MADR is an important agency in the three subsystems, however their participation in programmatic and OSN systems is less important than industrial agencies like MICIT or MINTIC. It is important to see the presence of the president Juan Manuel Santos in the discussion about innovation. The centrality measure in terms of closeness implies the tendency to reach the president and also their capacity to discuss several messages about innovation in Colombia. Network density in PN is higher than programmatic system, and also that OSN. This means that the ability to relate agencies and actors became lower between policy definition, execution and public discussion. Then, the coordination diluted despite the main role of Colciencias, that seems to keep the leadership to manage the NIS in Colombia.

Table 1. Compared networks

	Degree	Closeness	Betweenness	Density
Regulatory system	17,526	0,507	0,049	0,562
Colciencias	110	1.00	0.0876	
MEN	110	1.00	0.0876	
DNP	110	1.00	0.0876	
MADR	110	1.00	0.0876	
MCIT	96	0.9016	0.0517	
MINTIC	85	0.8209	0.0212	
SENA	85	0.8209	0.0212	
Programmatic system	7,600	N/A	0,419	0,135
Colciencias	18	0,62	0,459	
DNP	4	0,39		
MEN				
MADR	4	0,325	0,002	
MCIT	8	0,48	0,131	

SENA	13	0,59	0,290	
MINTIC	2	0,27	0,000	
MPS				
INNPULSA - UDI	4	0,39	0,1083	
OSN system	23,439	N/A	0,044	0,013
Colciencias	50	0,285	0,045	
DNP	19	0,234	0,005	
MEN	13	0,219	0,007	
MCIT	7	0,170	0,004	
Juan Manuel Santos	24	0,272	0,000	
MINTIC	16	0,234	0,008	
SENA	16	0,202	0,009	
MTIC	25	0,245	0,007	
MPS	3	0,000	0,000	
INNPULSA – UDI	24	0,245	0,002	

Source: Law 1286/2009; Law 1530/2012 and Decree 1500/2012, Colciencias Innovation Database 2010-2014 and Twitter Database 2010-2014.

Conclusion

This research presents the landscape of inter-agency collaboration in the innovation system in Colombia. We use a multilevel approach to determine the public agencies interactions since 2009, year in which the Law of the National System of Science, Technology and Innovation was re-launched in Colombia with a specific mandate to assure resources from royalties that increased the public investment in R&D to 2014.

The literature that evaluates the Colombian policy on science, technology and innovation reveals problems of efficiency. As Correa et al (2014, p. 275) points out: "The major problems in the system are: slow dissemination of information, missing deadlines, coherence of the guidelines given to the participants of the system and the need for steady and transparent regulatory system. In the case of Colombia the deadlines for receiving public bids are constantly delayed by the executing institutions with constant changes in the rules of the games and constant failure of the information system". This conclusion reflects the lack of coordination among several agencies that has the mandatory duty to define policies, execute programs and lead the public discussions on innovation.

Also, the book edited by Salazar and Fog (2013) presents the history of Colciencias empathizing their troubles, especially in actual years, to become a legitimate agency that can steer the national policy on innovation. This work reveals several shortcomings in the capacity of Colciencias to lead the NIS and its capacity to steer innovation policy is challenged by other agencies that have a stake on innovation and entrepreneurship like Innpulsa and Bancoldex from MINCIT. However, the results show that Colciencias still acts as the main agency in the promotion of innovation in Colombia.

Except for Colciencias, the main agencies of the Colombian Innovation system, as portrayed by the regulatory system, do not appear to play important roles at either the programmatic or the social network systems of innovation. This situation reflects the difficulty in Colombia to transition from a PN towards a GN. In fact, the three regulations on innovation create institutional mismatches, making inter-agency coordination difficult if not impossible, where Colciencias appears to be in practice the only organization capable of playing a coordinating role within the three systems studied.

Key agencies upon which the whole system is based, and for which a central role is expected within the three systems, appear in fact to be unable to coordinate and participate in a systemic

way, where competition, duplication of efforts and isolation are the guiding rules and result. One of the consequences (and potential contribution to) of the aforementioned lack of coordination within the NIS despite Colciencias' efforts, has to do with the recent decision made by the MADR to create an independent system of innovation for its ow sector, ignoring any possibilities to interact with Colciencias and other agencies and even disappearing from the regulatory, programmatic and social networks as the ones discussed here.

Another example refers to MINCIT. In 2014 the subdirector of the ministry declared that the NIS "no lo debería manejar Colciencias. Lo dice la OCDE: los esquemas de innovación deben ser administrados por la entidad de industria o desarrollo empresarial del país" (Revista Dinero Emprende, 2014a, p. 11). Moreover, the MINCIT created in 2012 Innpulsa to execute investments on innovation, while Colciencias has been reducing the budget to promote STI. These agencies tend to compete rather than to collaborate, as the Colciencias director of technology and innovation policy declared in an interview at the OCAD, Colciencias had their hands tied in the discussions on projects presented by governors. Revista Dinero (2014) described the collision between governors, universities and Colciencias. Governors said that Colciencias does not allow the approval of research projects until they decide the organizations and procedures to govern the resources, delaying the decision making.

In 2015 Colombia call for the integration of the three regulations studied here in the Law 1753- the National Developing Planning 2015-2018, claiming for a new model of governance that relate several agencies and actors that seek policies from research, innovation and competitiveness. However, the process need studies that reveal the shortcuts and problems in the cross-agency collaboration. The practical implication of this research is the description of PN and the evolution of GN in terms of the role of Colciencias, that is claiming to transform in a ministry that will be able to perform the coordination of policy definition, execution and discussion in the public realm of social networks.

The evolution of the system in terms of new officials ahead of these agencies could represent an opportunity to create new possibilities for coordination and collaboration between public agencies, with the involvement of several actors that seek to be heard and attended, like regional governors from Vichada or Choco (the regions that are less developed in the country). The construction of a NIS from the perspective of the PN and GNs is a bargaining process between several actors that interact in the definition, execution and accountability of policies in the regulatory, programmatic and governance systems. We believe that without the competitive regimen created by the regulatory system, more independence from normative proposals enacted by multilateral agencies like OCDE, and more inclusion of civil society, Colombia, running toward peace, could assume the control and accountability of their NIS leaded by Colciencias.

Acknowledgments

In memorial of José Luis Villaveces Cardoso (1945-2019).

References

Alcorta, L., & Peres, W. (1998). Innovation systems and technological specialization in Latin America and the Caribbean. Research Policy, 26(7), 857-881.

Arocena, R., and J. Sutz. 2002. Innovations Systems and Developing Countries. DRUID Working Paper No. 02-05. Äalborg, Denmark: Danish Research Unit for Industrial Dynamics, Äalborg University.

Audretsch, D. B. (2004). Sustaining Innovation and Growth: Public Policy Support for Entrepreneurship. Industry & Innovation, 11(3), 167–191.

Beer, S. (1984). The viable system model: Its provenance, development, methodology and pathology. Journal of the operational research society, 35(1), 7-25.

Blanco, I., Lowndes, V., & Pratchett, L. (2011). Policy networks and governance networks: Towards greater conceptual clarity. Political studies review, 9(3), 297-308.
Casas, R; Corona, JM y Rivera, R. (2014) "Políticas de ciencia, tecnología e innovación en América Latina: entre la competitividad y la inclusión social" en Kreimer, P; Vessuri, H; Velho, L y Arellano, A (Coord.) Perspectivas latinoamericanas en el estudio social de la ciencia, la tecnología y la sociedad (México: CYTED, ESOCITE y Foro Consultivo Científico y Tecnológico).
Chaminade, C, and Padilla, R. (2017). The challenge of alignment and barriers for the design and implementation of science, technology and innovation policies for innovation systems in developing countries. In S. Kuhlmann & G. Ordonez-Matamoros (Eds.), Research Handbook on Innovation Governance for Emerging Economies: Towards Better Models. Cheltenham, UK. and Northampton, MA. USA.: Edward Elgar
Correa, J. S., Tejada, M. A., Cayon Fallon, E., and Ordoñez, G. (2014). Science and Technology Policy in Colombia: A Comparative Review. European Journal of Scientific Research, 121(3), 267-285.
Daley, D. M. (2009). Interdisciplinary problems and agency boundaries: Exploring ffective cross-agency collaboration. Journal of Public Administration Research and Theory, 19(3), 477–493.
Deephouse, D. L., & Suchman, M. 2008. Legitimacy in organizational institutionalism. In R. Greenwood, C. Oliver, R. Suddaby, & K. Sahlin-Andersson (Eds.), The Sage handbook of organizational institutionalism: 49 –77. Thousand Oaks, CA: Sage.
Dutrénit, G, & Puchet, M. (2017). Tensions of Science, Technology and Innovation policy in Mexico: analytical models, institutional evolution, national capabilities and governance. In S. Kuhlmann & G. Ordonez-Matamoros (Eds.), Research Handbook on Innovation Governance for Emerging Economies: Towards Better Models. Cheltenham, UK. and Northampton, MA. USA.: Edward Elgar
Fawcett, P., & Daugbjerg, C. (2012). Explaining Governance Outcomes: Epistemology, Network Governance and Policy Network Analysis. Political Studies Review, 10(2), 195–207.
Fountain, JE (2002). "Toward a Theory of Federal Bureaucracy for the Twenty-First Century" Governance.com: Democracy in the Information Age. Ed. Elaine Ciulla Kamarck, Joseph S. Nye, Jr.. Washington, DC: Brookings Institution Press, 2002. 117-140.
Hudson, B., Hardy, B., Henwood, M., & Wistow, G. (1999). Inter-agency collaboration in the public sector What is the contrihution of theory and research? Public Management: An International Journal of Research and Theory, 1(2), 235–260.
Karo, E. (2010). Improving governance of science and innovation policies, or just bad policy emulation; the case of the estonian R&D system. Halduskultuur - Administrative Culture, 11(2), 174–201.
Karo, E. (2012). Modernizing governance of innovation policy through "decentralization": A new fashion or a threat to state capacities? Innovation: Management, Policy and Practice, 14(4), 495–509.
Kuhlmann, S. and Shapira, P. (2006) How is Innovation Influenced by Science and Technology Policy Governance? Transatlantic Comparisons. In: Hage, J. / Meeus, M. (eds.): Innovation, Science, and Institutional Change; A Research Handbook, Oxford, Oxford University Press.
Kuhlmann, Stefan, and Ordonez-Matamoros, H.G. (Eds.). (2017). Research Handbook on Innovation Governance for Emerging Economies: Towards Better Models. Cheltenham, U.K. and Northampton, MA. USA.: Edward Elgar
Laranja, M. (2012). Network governance of innovation policies: The Technological Plan in Portugal. Science and Public Policy, 39(5), 655–668.
Lindner, R. , Kuhlmann, S., Randles, S., Bedsted, B., Gorgoni, G., Griessler, E., . . . Mejlgaard, V. (2016). Navigating Towards Shared Responsibility in Research and Innovation. Approach, Process and Results of the Res-AGorA Project
Lips, A. M. B., O'Neill, R. R., & Eppel, E. A. (2011). Cross-agency collaboration in New Zealand: An empirical study of information sharing practices, enablers and barriers in managing for shared social outcomes.international Journal of public Administration, 34(4), 255-266.
Lundvall, B. Å., Chaminade, C., Joseph, K. J., & Vang, J. (Eds.). (2009). Handbook on innovation systems and developing countries: Building domestic capabilities in a global context. Cheltenham, England: Edward Elgar

Orozco, L_A; Chavarro, D_A; Olaya, D y Villaveces, J_L (2007) “Methodology for measuring the socio-economic impacts of biotechnology: a case study of potatoes in Colombia”. Research Evaluation. 16[2]: 107-122.
Revista Dinero Emprende (2014). Hay que creernos el cuento. Somos capaces. Revista Dinero Emprende. May 2014. Bogotá.
Revista Dinero (2014). Una verdad incómoda. April 2014. Bogotá. Available at: http://www.dinero.com/edicion-impresa/pais/articulo/regalias-para-ciencia-tecnologia/194705
Salazar, M y Fog, L (Ed) (2013). Colciencias cuarenta años: entre la legitimidad, la normatividad y la práctica. Bogotá: Observatorio Colombiano de Ciencia y Tecnología - OCyT.
Sánchez, JM and Pérez, P (2013). Desde sus inicios. El trasegar de Colciencias en la promoción de la innovación. In: Salazar, M y Fog, L (Ed). Colciencias cuarenta años: entre la legitimidad, la normatividad y la práctica. Bogotá: Observatorio Colombiano de Ciencia y Tecnología - OCyT.
Schepers, S. (2013). Governance and innovation policy failures in Europe. Journal of Public Affairs, 13(3), 345-351.
Scott, RW (2005). Institutions and organizations. (2ed). Sage Thousand Oak. CA.
Yun, S. J., Ku, D., & Han, J. Y. (2014). Climate policy networks in South Korea: alliances and conflicts. Climate policy, 14(2), 283-301.

What makes some scientific findings more certain than others? A study of citing sentences for low-hedged papers

Henry Small

hsmalll@mapofscience.com
SciTech Strategies Inc., Bala Cynwyd, PA 19004 (USA)

Abstract
The citing sentences for low-hedged and highly hedged papers were compared in terms of word usage with the aim of contrasting the epistemic styles associated with high and low certainty science, and determine the criteria used for creating credible findings. It was found that low-hedged, or high certainty, papers were associated with action verbs denoting the application of methods and acquisition of data, and words specifically denoting quantitative methods. High-hedged, or low certainty, papers, on the other hand, were associated with words denoting interpretation and justification of ideas. A linear regression was able to account for 70% of the variance in the hedging rate using the word variables plus a consensus measure based on the degree of similarity of citing sentences for a given paper. A co-occurrence map of words associated with high and low hedged citing sentences showed a separation of method and interpretation words.

Introduction

Citing sentences, sentences in papers containing references to earlier papers, sometimes called citances, can be used to study citing authors' disposition towards the earlier papers. This is especially powerful when multiple citing sentences can be aggregated for a specific highly cited paper. For example, if citing sentences for a given cited paper contain hedging terms, then we can infer that citing authors consider that work as uncertain in some respect. Conversely, if the citing sentences for a paper are infrequently hedged or not hedged at all, then we can infer that authors regard that paper, or its findings, with some degree of certainty. The "certainty" or "uncertainty" of scientific findings is a question of fundamental importance in the history and philosophy of science (Rescher, 2006), science policy (Oreskes & Conway, 2010), and everyday life. The aim of this paper is to study citing sentences associated with low-hedged, highly cited papers to identify the characteristics of "certainty" in science from the point of view of the citing authors. This will allow us to examine the strategies and criteria scientists employ when creating valid or certain scientific findings. This might be termed "reverse epistemology", where we use text describing actual scientific practice to back into the philosophical question of what determines certainty in science. This research can also be considered a contribution to the general area of "discourse epistemetrics" (Demarest & Sugimoto, 2015).

Method

An important underpinning of this research is previous work on hedging in scientific texts (Chen & Song, 2018; Hyland, 2009). Hedging terms have been observed to be overrepresented in citing sentences (DiMarco, Kroon & Mercer, 2005). For our purposes we do not need to identify every possible form of hedging in scientific texts, only the most prevalent forms. A list of 50 common hedging terms was compiled from several sources (Small & Klavans, 2011), and the frequency with which these terms appeared in citing sentences in our data source was determined. The three most prominent hedging terms were "may", "could" and "might". The percentage of citing

sentences that contain one or more of these three words was computed for a set of highly cited papers. This provides an approximate hedging rate for each cited paper that can be used as a dependent variable.

A previous study that combined citation analysis with textual analysis of citing sentences looked at highly cited method papers. It was found that the most prominent words in the citing sentences denoted utility, such as "using" and "used" (Small, 2018). It was also found that method papers were not hedged as often as non-methods and that the higher the number of citations a paper received, whether it was a method or not, the lower its rate of being hedged. Non-methods included, for example, discovery and review papers.

The PubMed Central open access subset provides access to the full text of biomedical papers with XML markup. This allows us to connect hedging in citing sentences and the specific paper cited in the text by using reference tags that connect the in-text references to the reference lists at the end of papers. Only papers with a PubMed ID were used in this analysis.

The strategy was to define samples of highly hedged papers and infrequently hedged papers from the set of 1,000 biomedical papers in the PubMed Central database having the highest number of citances. A "citance count" is like a traditional "citation count" but uses the total number of mentions in full-texts of citing papers rather than counting each reference one time. In the low and high hedged samples from the top 1,000 list, the hedging rate ranged from 0 to a maximum rate of 27 percent. There were 130,203 citing sentences in the low-hedged sample and 89,237 citing sentences in the high-hedged sample. The average citance count of the papers in the highly hedged sample was 929 citances per paper, compared to a mean of 1,463 for the low-hedged sample, which is consistent with the prior finding that more highly cited papers are less hedged than less cited papers.

The software package WordSmith Tools (Scott, 2004) was used to compute word frequencies for low versus high-hedged samples. This software tool computes log likelihood statistics for each word in one sample using a second sample as a comparison or baseline set. All words in the citing sentences were counted. Most words are, of course, technical in nature, but the general words such as "using", "performed" and "data" are of special interest since they tell us about the nature of the scientific activity being described. In addition, the rates of occurrence of these words in the citances of the highly cited papers were computed, and a regression was carried out to see to what extent the hedging rate could be predicted. Finally, a mapping based on the co-occurrence of the words in citances was performed to see if epistemic styles could be discerned.

Results

The low-hedged citing sentences were input to WordSmith with the high-hedged citing sentences as the baseline to see if differences in word usage rates suggest differences in certainty. The inverse comparison, the high-hedged sample with respect to the low-hedged sample gives the view from uncertainty. A selection of terms having significant log likelihoods for each comparison are given in Table1 which is designed to highlight the differences between low and high hedged samples.

Table 1 reveals that words from low-hedged citances are dominated by action verbs ("determined", "performed" and "calculated"), and the acquisition of data, as exemplified by the following citance: "Relative **quantification** of gene expression was **calculated** using the REST **software** tool." (Bolded words are keywords in the analysis.) A set of 17 action verbs was able to retrieve 37 percent of the low-hedged citances, meaning that each sentence contained one or more of these action words, but only 7 percent of high-hedged citances was retrieved. High-hedged citance words, on the other hand, are interpretative and tentative ("may", "association", "suggested") and are concerned with justification ("evidence", "important", "explain"), for example: "Increasing **evidence suggests** that inflammatory cells are an essential component of the tumor microenvironment . . . and progression and **may** be associated with systemic inflammation." A set of 18 of these interpretative words (including "evidence", "important", "predicted" and "explained") retrieved 35 percent of high-hedged citances compared to 12 percent of low-hedged citances, not as striking a difference as was seen for action words but still significant ($p < .05$ with a chi-square test for difference in proportions). In summary, the contrast between low and high hedging is between action and interpretation.

Table 1: General keywords with significant log likelihoods (LL) for low-hedged papers (left column) and high-hedged papers (right column).

Low-hedged papers	LL	High-hedged papers	LL
using	36,090	may	5,849
method	13,883	guidelines	5,146
software	10,283	evidence	3,691
determined	6,050	association	2,200
performed	5,174	criteria	1,687
likelihood	3,360	selection	1,581
standard	3,117	recommended	1,367
calculated	3,101	suggested	1,141
estimated	1,829	recent	1,122
analyzed	1,633	important	1,072
measured	1,308	proposed	962
quantified	1,167	demonstrated	904
normalized	1,101	prediction	763
parameter	861	explain	664

To get more detail on the phenomenon of low-hedging we split the sample into method and non-method papers. Citances for each subset are then analyzed using the high-hedged corpus as the baseline set. Table 2 contains a selection of words with significant log likelihoods. Again, we see action verbs, such as "performed", "determined" and "analyzed", associated with low-hedged method papers, and a focus on operations undertaken to generate data. Words, such as "cause", "cases" and "estimated" are, on the other hand, associated with low-hedged, non-method papers suggesting that data is being deployed rather than generated, to back up conclusions or draw

inferences. For example: "The hypertension **prevalence** is increasing world-wide and it is **predicted** that 60% of the population will suffer from hypertension in 2025."

Table 2: General keywords with significant log likelihoods (LL) for low-hedged method (left column) and non-method papers (right column). A "†" indicates a quantitative word.

Low-hedged method papers	LL	Low-hedged non-method papers	LL
using	49,886	cause	3,018
method	17,642	cases†	2,548
performed	7,941	incidence†	1,528
determined	7,245	estimated†	1,298
analyzed	5,149	rate†	1,172
calculated†	4,508	database	1,135
data†	3,417	median†	1,056
measured†	2,097	prevalence†	738
described	1,636	percentile†	459
values†	1,514	dataset†	452
obtained	1,473	obtained	327
quantified†	1,343	ranks†	305
normalized†	1,307	discovered	303

Many of the words in Table 2 for both method and non-method papers describe quantitative operations, such as "calculated", "data", "measured", "cases", "estimated", "dataset", etc. A list of 15 quantitative words was able to retrieve 35 percent of the low-hedged citances but only 22 percent the high-hedged sentences ($p < .05$). This confirms the quantitative orientation of the low-hedged sample, and we can conclude that quantitative data are important for establishing certainty in biomedical research.

Having identified prominent words used by citing authors to characterize high and low-hedged papers, the next question is to what extent they can predict the hedging rate. Linear regression can be used to model the hedging rate as a dependent variable using specific citing sentence word percentages combined with other variables. The log transform of the hedging rate was used to make the relationship of hedging to other variables more linear.

Stephen Cole had suggested that consensus is an important determinant of citations (Cole, 1992). A measure of consensus was proposed as the degree of similarity of citing sentences for a given cited paper (Small, 2018). A consensus metric was implemented by computing the average cosine similarity of each citing sentence with the cumulation of all citing sentences for a given cited paper. The metric was computed for each of the papers in the high and low hedged samples. Following Cole, consensus should be positively correlated with certainty. Another variable is the age of the cited paper computed as the difference between 2017, the final year of the PubMed Central data under analysis, and the publication year of the paper. Older papers were expected to reflect more settled science and have higher certainty. A final variable was whether the paper was categorized

as a method (1 for method and zero for non-method) since methods were found to have higher certainty in a prior study (Small, 2018).

The following independent variables were used in the regression: age of paper, consensus, and several word rates obtained by searching for word stems or word combinations in citances for individual papers. All words included had been found in the WordSmith analyses as keywords with statistically significant log likelihoods. Multiple regression was carried out in "R" (R Core Team, 2017), and gave the results in Table 3 which shows nine statistically significant variables after a stepwise regression (Teetor, 2011).

Table 3: Linear regression to predict hedging rate

log hedge ~ age + consensus + using/used + cause + cases + described + measured + obtained + performed

Residuals:

Min	1Q	Median	3Q	Max
-3.5297	-0.3044	0.0233	0.3359	2.6803

Variable	Est. Coefficient	Std. Error	t value	Pr(>|t|)
(intercept)	3.275087	0.100843	32.477	< 2e-16 ***
age	-0.008094	0.002305	-3.512	0.000466 ***
consensus	-3.529769	0.335411	-10.524	< 2e-16 ***
using/used	-0.023246	0.001260	-18.448	< 2e-16 ***
cause	-0.026470	0.004933	-5.366	1.00e-07 ***
cases	-0.029674	0.008231	-3.605	0.000328 ***
described	-0.032051	0.004886	-6.559	8.72e-11 ***
measured	-0.032415	0.008393	-3.862	0.000120 ***
obtained	-0.067799	0.012062	-5.621	2.47e-08 ***
performed	-0.026523	0.005086	-5.214	2.25e-07 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5359 on 987 degrees of freedom
Multiple R-squared: 0.7357, Adjusted R-squared: 0.7333
F-statistic: 305.3 on 9 and 987 DF, p-value: < 2.2e-16

The model is a relatively good fit to the data, explaining over 70 percent of the variance. The regression coefficients are negative except the intercept indicating that the higher the value of the variable the lower the hedging, and the more certain the finding. The best predicting words are a mix of action and data words.

The log likelihood analysis has identified various types of words associated with high and low hedged citances which we have described as action, interpretative, or quantitative words. Sometimes these words occur jointly in a citance, and thus it was of interest to see whether a co-

occurrence mapping would reveal any structure. Figure 1 shows one such mapping based on 378 words that co-occur in over 1 million citances for the top 1,000 papers. Word co-occurrence was normalized by the cosine formula and edges with a cosine of 1 or more were mapped using VOSviewer (Van Eck & Waltman, 2010). The words appear to form groups with method/action words on the right and interpretative/evidential words on the left. A region on the upper right deals with quantitative words.

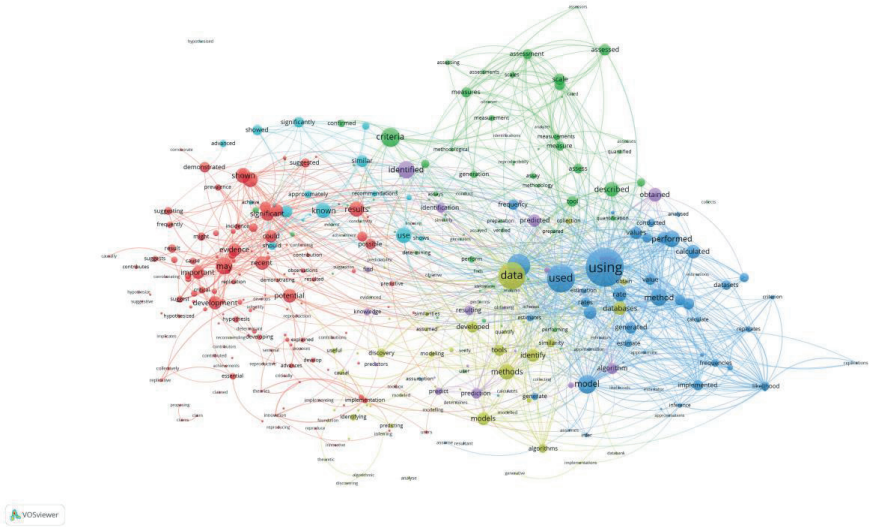


Figure 1. Co-occurrence map of general keywords showing groupings for epistemic styles.

Conclusions

The assumption of this paper is that we can ascertain the criteria applied to establish the certainty or validity of scientific results by analyzing the citing behavior of scientists. This might be considered reverse epistemology because, rather than applying some philosophical definition of certainty or validity, we are attempting to infer such a definition empirically. We first identify papers that are of relatively high certainty as indicated by the absence of hedging in citing sentences, and then further analyze the citing sentence words to ascertain the dominant epistemic style. The same approach is applied to the uncertain or highly-hedged papers. The difference between low and high hedged papers appears to lie on the dimensions of quantitative/active versus qualitative/interpretative approaches. We have found that low-hedged papers are split between efforts to generate data and efforts to draw conclusions from data. The higher rate of usage of quantitative key words in low versus high hedged citing sentences offers further evidence of this. A regression analysis, in addition to confirming the importance of certain key words in predicting hedging rate, also found consensus as an important predictor: the higher the consensus, the lower

the hedging and higher the certainty. A mapping of selected words associated with high and low hedged citances confirms a bifurcation of epistemic styles.

Some obvious ways the present research could be improved are the use of a wider set of hedging terms, expanding the sample sizes of papers and citing sentences, and extension to other disciplines. Another avenue of research that might be amenable to "applied epistemology" is exploring the criteria for confirmation and disconfirmation in science.

References

Chen, C. & Song, M. (2018). Representing scientific knowledge: the role of uncertainty. London, UK: Springer.

Cole, S. (1992). *Making science: between nature and society*. Cambridge, Mass.: Harvard University Press.

Demarest, B. & Sugimoto, C.R. (2015). Argue, observe, assess: measuring disciplinary identities and difference through socio-epistemic discourse. *Journal of the American Society for Information Science*, 45, 12-19.

DiMarco, C., Kroon, F.W., & Mercer, R.E. (2006). Using hedges to classify citations in scientific articles. In J. Shanahan, Y. Qu & J. Wiebe (Eds.), *Computing attitude and affect in text: theory and applications, vol. 20. The Information Retrieval Series.* (pp. 247-263). Netherlands: Springer.

Hyland, K. (2009). *Academic discourse: English in a global context*. London.: Continuum.

Oreskes, N. & Conway, E.M. (2010). *Merchants of doubt*. New York: Bloomsbury Press.

R Core Team (2017). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Rescher, N. (2006). *Epistemetrics*. New York: Cambridge University Press.

Scott, M. (2004). *WordSmith tools version 4*. Oxford: Oxford University Press.

Small, H. (2018). Characterizing highly cited method and non-method papers using citation contexts: the role of uncertainty. *Journal of Informetrics*, 12(2), 461-480.

Small, H. & Klavans, R. (2011). Identifying scientific breakthroughs by combining co-citation analysis and citation context. *Proceedings of the 13th International Conference of the International Society for Scientometrics and Informetrics,* Durban, South Africa.

Teetor, P. (2011). *R Cookbook.* Sebastopol, CA: O'Reilly Media, Inc.

Van Eck, N.J., & Waltman, L. (2010). Software survey: VOSviewer, a computer program for bibliometric mapping. Scientometrics, 84(2), 523–538.

A multidimensional perspective on the citation impact of scientific publications

Yi Bu[1], Ludo Waltman[2], and Yong Huang[3]

[1] *buyi@iu.edu*
Center for Complex Networks and Systems Research, School of Informatics, Computing, and Engineering, Indiana University, Bloomington, IN (U.S.A.)

[2] *waltmanlr@cwts.leidenuniv.nl*
Centre for Science and Technology Studies, Leiden University, Leiden (The Netherlands)

[3] *yonghuang1991@whu.edu.cn*
School of Information Management, Wuhan University, Wuhan, Hubei (China)

Abstract

The citation impact of scientific publications is usually seen as a one-dimensional concept. We introduce a three-dimensional perspective on the citation impact of publications. In addition to the level of citation impact, quantified by the number of citations received by a publication, we also conceptualize and operationalize the depth and dependence of citation impact. This enables us to make a distinction between publications that have a deep impact concentrated in one specific field of research and publications that have a broad impact scattered over different research fields. It also allows us to distinguish between publications that are strongly dependent on earlier work and publications that make a more independent scientific contribution. We present a large-scale empirical analysis of the level, depth, and dependence of the citation impact of publications. In addition, we report a case study focusing on publications in the field of scientometrics. Our three-dimensional citation impact framework provides a more detailed understanding of the citation impact of a publication than a traditional one-dimensional perspective.

Introduction

Measuring the citation impact of scientific publications is an important topic in bibliometric and scientometric research. Many different citation impact indicators, calculated based on the citations received by a publication, have been proposed (Waltman, 2016). The most basic citation impact indicator is the raw citation count of a publication. Although this indicator is easy to calculate, it has often been criticized and many alternatives have been proposed.

Normalization is a commonly used approach to construct more sophisticated citation impact indicators (Waltman & Van Eck, 2018). Several attributes of a publication have been used for normalization, in particular a publication's scientific field and its age. Another prominent line of research on citation impact indicators focuses on PageRank-inspired approaches (Waltman & Yan, 2014). For instance, Chen, Xie, Maslov, and Redner (2007) proposed a PageRank approach for quantifying the citation impact of a publication. Attributes derived from the full text of citing publications, such as the number of times a publication is cited in the full text of a citing publication and the location in the full text where the publication is cited, have also been suggested as useful features for constructing citation impact indicators (e.g., Ding, Liu, Guo, & Cronin, 2013; Wan & Liu, 2014; Zhu, Turney, Lemire, & Vellino, 2015).

The approaches discussed above still regard the citation impact of a publication as a one-dimensional concept. In this paper, we propose a multidimensional perspective on the citation impact of a publication. We argue that, in addition to the level of citation impact, there are other interesting aspects of the citation impact of a publication that can be derived from the citations received by a publication.

To illustrate this, consider two publications, A and B. As shown in Figure 1, these publications have both received five citations. If we just count the citations received by the two

publications, the publications will be considered to have the same citation impact. However, the publications citing A also cite each other and therefore seem to be closely related, while the publications citing B do not cite each other and therefore seem to be quite unrelated from each other. Hence, A and B have the same level of citation impact, but A seems to have a relatively deep citation impact in a narrow research area, while B seems to have a broader citation impact. To distinguish between the different ways in which A and B have an impact on other publications, we propose an approach for quantifying the depth of the citation impact of a publication.

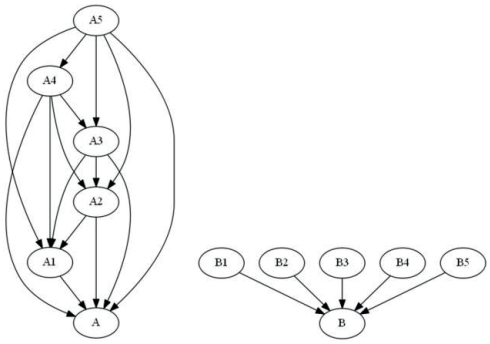


Figure 1. Deep vs. broad citation impact. Nodes represent publications and edges represent citation relations. Publications *A* and *B* have both received five citations. All publications citing *A* also cite each other, which is not the case for publications citing *B*. Therefore *A* has a deep citation impact, while *B* has a broad citation impact.

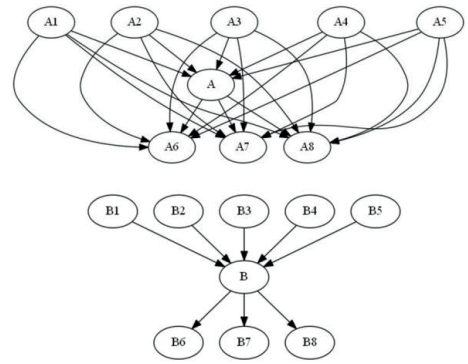


Figure 2. Dependent vs. independent citation impact. Nodes represent publications and edges represent citation relations. Publications *A* and *B* have both received five citations, and they both have three references. All publications citing *A* also cite each of the references of *A*, which is not the case for publications citing *B*. Therefore *A* has a citation impact that is strongly dependent on earlier publications, while *B* has an independent citation impact.

We are also interested in the dependence of a publication's citation impact on earlier publications. For instance, consider two publications, A and B. As shown in Figure 2, these publications have both received five citations, and they both have three references. All publications citing A also cite each of A's references, while the publications citing B do not cite B's references. Hence, the citation impact of A seems to depend strongly on earlier publications, namely the ones cited by A. It is likely that A is a follow-up study of these earlier publications. On the other hand, B seems to have a much more independent citation impact, since publications citing B do not cite the references of B.

In this paper, we propose to conceptualize and operationalize the citation impact of a publication in a three-dimensional framework that focuses on the level, depth, and dependence of citation impact. In a traditional one-dimensional perspective on citation impact, only the level of citation impact is considered. Beyond the level of citation impact, no insights are obtained into the way in which a publication has an impact on other publications. By introducing the concepts of depth and dependence, our proposed framework aims to offer a more detailed understanding of the citation impact of a publication.

The idea of analyzing citation relations between publications that cite a focal publication is not new. Clough, Gollings, Loach, and Evans (2015) compared the number of citations given to a publication in a citation network with the number of citations given to the same publication

in the transitive reduction of the citation network. According to Clough et al., the transitive reduction can be used to get "an indication that results in a paper were used across a wide number of fields". Huang, Bu, Ding, and Lu (2018a, 2018b) analyzed so-called citing cascades, defined as the citation network of a focal publication and its citing publications. In particular, they studied citation relations between citing publications. The citation impact framework proposed in the current paper partly builds on the ideas explored by Huang et al. The notion of dependence introduced in our citation impact framework is also related to the concepts of development and disruption recently proposed by Funk and Owen-Smith (2016) as well as Wu, Wang, and Evans (2019).

Level, depth, and dependence of the citation impact of a publication

In this section, we present our three-dimensional framework for characterizing the citation impact of publications. We first discuss the level of citation impact, followed by the depth of citation impact and finally the dependence of citation impact. Throughout this section, we focus on a publication X that has m references and that has been cited by n other publications, denoted by $Y_1, Y_2, \ldots, Y_n$.

Level of citation impact

The level of citation impact of a publication reflects how much impact the publication has had on other publications. We operationalize this by the number of citations a publication has received. Hence, the level of citation impact of publication X, denoted by level_X, equals

$$\text{level}_X = n. \tag{1}$$

The larger the number of citations received by a publication, the higher the level of citation impact of the publication. The level of citation impact coincides with the traditional way in which the citation impact of a publication is conceptualized and operationalized.

Depth of citation impact

To understand the notion of the depth of the citation impact of a publication, we consider an example involving two publications, A and B. These publications have received the same number of citations, and they therefore have the same level of citation impact. However, A and B differ in how they have an impact on other publications. Let's first consider A. Suppose A introduces an innovative new idea in a certain research field. Many publications in this field start to build on this idea. These publications all cite A and many of them also cite each other. On the other hand, outside the research field of A, little attention is paid to the idea introduced in A and relatively few citations are made to A. Let's now consider B. Suppose B introduces a new software tool for carrying out certain statistical analyses. The tool turns out to be useful in many different research fields. In all these fields, publications that use the tool cite B. However, apart from the fact that they use the tool introduced in B, these publications have very little in common. They all deal with different research questions. In general, publications citing B therefore do not cite each other.

In this example, it is clear that A and B have an impact on other publications in very different ways. We consider A to have a deep citation impact, while we consider B to have a broad citation impact. Hence, if the impact of a publication is strongly concentrated within a single research field, the publication has a deep citation impact. On the other hand, if the impact of a publication is scattered over many different research fields, the publication has a broad citation impact. We treat depth and breadth of citation impact as opposite concepts. Consequently, a high depth implies a low breadth, and the other way around. In our operationalization, we focus on the depth of citation impact.

The depth of the citation impact of publication X, denoted by depth_X, is given by

$$\text{depth}_X = \frac{1}{n}\sum_{i=1}^{n} CC_{X,i}, \tag{2}$$

where $CC_{X,i}$ denotes the number of co-citation links between X and Y_i, that is, the number of publications that cite both X and Y_i. In other words, the depth of the citation impact of publication X equals the average number of co-citation links between publications citing X and X itself. For instance, in Figure 1, the depth of the citation impact of A and B equals 10 / 5 = 2 and 0 / 5 = 0, respectively. This means that A has a deep citation impact, while B has a broad citation impact.

It follows from (2) that $\text{depth}_X \in [0, \frac{n-1}{2}]$. Hence, the upper bound of depth_X is determined by the number of citations X has received, suggesting there may be a positive correlation between depth_X and level_X. If $\text{depth}_X = 0$, there are no citation relations between publications that cite X. Publications citing X then do not seem to be related to each other and X has a broad citation impact. On the other hand, if $\text{depth}_X = \frac{n-1}{2}$, all publications that cite X are connected to each other by citation relations. In other words, publications citing X are strongly related and X has a deep citation impact. We note that depth_X is undefined if X has not been cited (i.e., $n = 0$).

We do not intend to make a normative judgment by quantifying the depth of the citation impact of a publication. From our point of view, a deep citation impact is not necessarily 'better' than a broad citation impact, or the other way around. However, we do believe that the depth vs. breadth distinction is useful to get a more detailed understanding of the way in which a publication has an impact on other publications.

Dependence of citation impact

Finally, we introduce the notion of the dependence of the citation impact of a publication. Publications may have a similar level and a similar depth of citation impact, but nevertheless there may be an important difference in how they have an impact on other publications. Some publications may have an impact by building on earlier publications and by contributing new scientific knowledge in a cumulative way. It is likely that these publications will usually be cited together with the publications on which they build. We consider the citation impact of these publications to have a high dependence. Other publications may have an impact without relying strongly on earlier publications. These publications may introduce new ideas that have been developed relatively independently from earlier literature. These publications will not be cited together with other publications on which they have a strong dependence. We consider these publications to have a relatively independent citation impact.

Our operationalization of the dependence of the citation impact of a publication mirrors the operationalization of the depth of citation impact provided above, with co-citation links being replaced by bibliographic coupling links. The dependence of the citation impact of publication X, denoted by dependence_X, is given by

$$\text{dependence}_X = \frac{1}{n}\sum_{i=1}^{n} BC_{X,i}, \tag{3}$$

where $BC_{X,i}$ denotes the number of bibliographic coupling links between X and Y_i, that is, the number of references that X and Y_i have in common. In other words, the dependence of the citation impact of X equals the average number of bibliographic coupling links between publications citing X and X itself. For instance, in Figure 2, the dependence of the citation impact of A and B equals 15 / 5 = 3 and 0 / 5 = 0, respectively.

Eq. (3) implies that $\text{dependence}_X \in [0, m]$. Hence, the upper bound of dependence_X is given by the number of references of X. This suggests that publications with longer reference lists, such as review articles, may tend to have a higher dependence. If $\text{dependence}_X = 0$, none of the publications citing X cites any of the references of X. There then seem to be no common dependencies of X and publications citing X on earlier literature, and X is therefore considered to have an independent citation impact. On the other hand, if $\text{dependence}_X = m$, each of the publications citing X also cites all references of X. This indicates that the citation impact of X depends strongly on the references of X. We note that dependence_X is undefined if X has not been cited (i.e., $n = 0$).

Empirical analysis

Data

Our empirical analysis was carried out using data extracted from the in-house version of the Web of Science (WoS) database available at the Centre for Science and Technology Studies (CWTS) at Leiden University. We made use of the Science Citation Index Expanded, the Social Sciences Citation Index, and the Arts & Humanities Citation Index. The data covers nearly 50 million publications that appeared between 1980 and 2017 and over 720 million citation relations between these publications. The analysis focuses on highly cited publications in the period 2000–2017. We defined a highly cited publication as a publication that has received at least 100 citations. We did not impose any document type restrictions. In total, 637,237 highly cited publications in the period 2000–2017 were identified. We calculated the level, depth, and dependence of the citation impact of these publications. The calculation of dependence required the identification of bibliographic coupling links. A bibliographic coupling link was identified between two publications if there is a third publication that is included in our data and that is cited by the other two publications. If two publications both cite the same publication but this publication is not included in our data, this did not result in the identification of a bibliographic coupling link.[1]

Overview

Figure 3 shows the cumulative distribution functions (CDFs) of the level, depth, and dependence of the citation impact of our 637,237 highly cited publications. The well-known skewness of citation distributions is reflected in the CDF of the level of citation impact. The CDFs of the depth and dependence of citation impact are surprisingly similar. The mean and median depth equal 2.8 and 2.4, respectively. The maximum depth equals 25.3. The mean and median dependence equal 2.7 and 2.3, respectively, while the maximum dependence equals 29.5.

Table 1 reports the Pearson and Spearman correlations between the level, depth, and dependence of the citation impact of our highly cited publications. The Pearson and Spearman correlations yield similar results. As expected, there is a positive correlation between level and depth, although this correlation is weak. Depth and dependence are also positively correlated. Despite the positive correlations between level and depth and between depth and dependence, there is a weak negative correlation between level and dependence.[2]

Figure 4 provides more detailed insight into the relations between the level, depth, and dependence of citation impact. The figure shows CDFs of depth and dependence for different levels of citation impact. In addition, it shows CDFs of dependence for different values of depth.

[1] This is the reason why our analysis focuses on highly cited publications in the period 2000–2017 and why highly cited publications in the period 1980–1999 are not considered. The calculation of dependence for older publications is affected by the fact that many references in these publications point to literature that appeared before 1980 and that is not included in our data.

[2] As expected, dependence is positively correlated with the number of references of a publication. This is a quite strong correlation ($r = 0.44$). The number of references has a weak positive correlation with level and a weak negative correlation with depth.

The results are in agreement with the correlations reported in Table 1. On average, the higher the level of citation impact, the higher the depth and the lower the dependence of citation impact. In addition, the higher the depth of citation impact, the higher the dependence. Figure 4 clearly shows how the most highly cited publications are characterized by a high depth and a low dependence. Overall, however, depth and dependence are positively correlated, and the publications with the highest depth therefore also tend to have a high dependence.

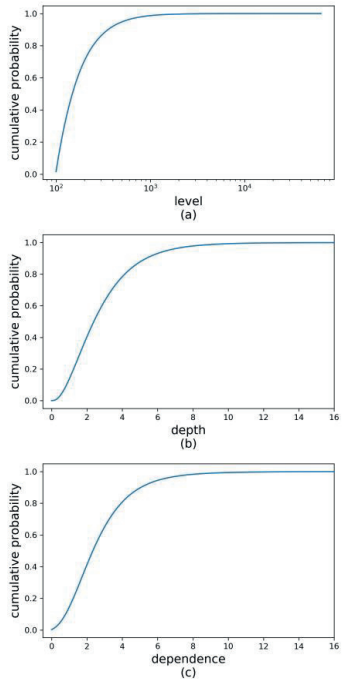


Figure 3. Cumulative distribution functions of the level, depth, and dependence of the citation impact of highly cited publications.

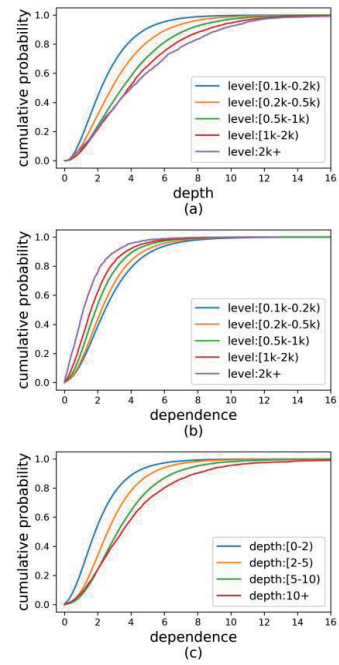


Figure 4. Cumulative distribution functions showing the relations between the level, depth, and dependence of the citation impact of highly cited publications.

We speculate that our findings may be explained by distinguishing between on the one hand publications contributing to research areas that are of a strongly cumulative nature and on the other hand publications making a more independent scientific contribution. In general, the former publications can be expected to have a relatively high depth and dependence, while the latter publications can be expected to have a lower depth and dependence. This would explain the positive correlation between depth and dependence. Cumulative research areas may start from highly influential "breakthrough" publications. These publications have a high level of citation impact and a high depth. However, because they are at the start of a new research area, they can be expected to have a low dependence. This would explain why the most highly cited publications tend to have a high depth and a low dependence.

Table 1. Pearson (upper-right triangle) and Spearman (bottom-left triangle) correlations between the level, depth, and dependence of the citation impact of highly cited publications.

	Level	*Depth*	*Dependence*
Level	1.00	0.15	-0.08
Depth	0.21	1.00	0.29
Dependence	-0.09	0.32	1.00

Table 2. Number of highly cited publications per discipline and the mean,

median, and maximum level, depth, and dependence of the citation impact of these publications.

	BHS	*LES*	*MCS*	*PSE*	*SSH*
No. of pub.	332,143	84,775	12,573	168,617	32,266
Mean level	217.7	201.6	209.6	218.6	206.6
Median level	151	144	144	149	150
Max. level	64,246	27,889	13,709	64,584	14,229
Mean depth	2.7	3.0	2.2	3.3	2.1
Median depth	2.3	2.6	1.9	2.8	1.8
Max. depth	25.3	20.5	17.1	23.1	13.5
Mean dependence	2.7	2.7	1.5	2.9	2.2
Median dependence	2.4	2.3	1.2	2.4	1.9
Max. dependence	28.6	26.5	16.8	29.5	17.6

Disciplinary comparison

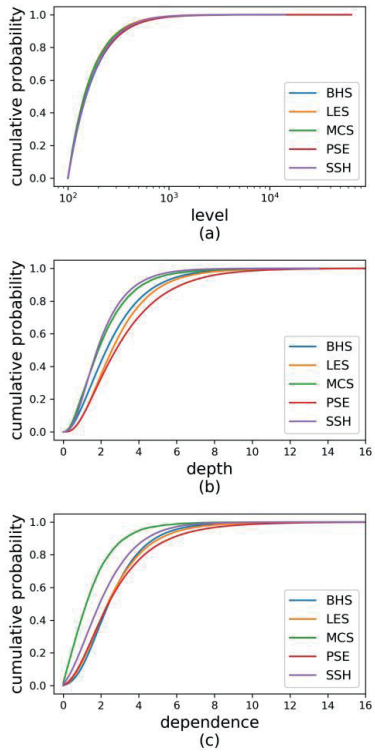


Figure 5. Cumulative distribution functions per discipline of the level, depth, and dependence of the citation impact of highly cited publications.

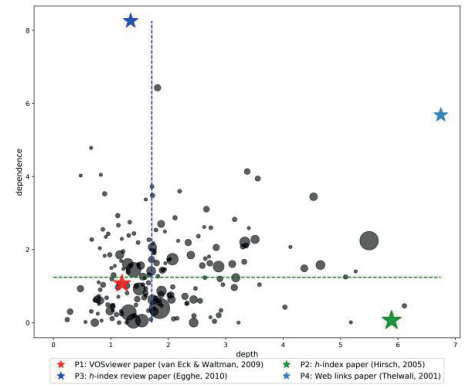


Figure 6. Scatter plot of the depth and dependence of the citation impact of 182 highly cited publications in the field of scientometrics. The dashed vertical and horizontal lines indicate the median depth and median dependence, respectively. The size of a point is proportional to the number of citations the corresponding publication has received. In each quadrant, one publication was selected, marked by a star.

Using the algorithmic methodology introduced by Waltman and Van Eck (2012), publications in our WoS database in the period 2000–2017 were clustered based on citation relations. Only publications of the document types 'article', 'letter', and 'review' were included in the clustering. 4,047 clusters of publications were obtained. Clusters are non-overlapping. Each publication belongs to only one cluster. The 4,047 clusters were grouped into the following five broad scientific disciplines: Biomedical and health sciences (BHS), Life and earth sciences (LES), Mathematics and computer science (MCS), Physical sciences and engineering (PSE),

and Social sciences and humanities (SSH). We now present a comparison of these five disciplines in terms of the level, depth, and dependence of the citation impact of publications. We again focus on highly cited publications, defined as publications that have received at least 100 citations.

Table 2 reports for each of the five disciplines the number of highly cited publications and the mean, median, and maximum level, depth, and dependence of the citation impact of these publications. The average level of citation impact is quite similar in the different disciplines. For the depth and dependence of citation impact, more substantial disciplinary differences can be observed. On average, the highest values of both depth and dependence can be found in BHS, LES, and PSE. This may reflect the cumulative nature of research in these disciplines. MCS and SSH are characterized by lower values of depth and dependence, suggesting that scientific contributions are more independent from each other in these disciplines. Especially the low values of dependence in MCS are noteworthy. However, the low values of dependence in MCS and, to a lesser extent, SSH may partly be an artifact of our data. When two publications both cite the same third publication, this is counted as a bibliographic coupling link only if the third publication is included in our data. Especially in the case of MCS and SSH, this is not always the case, since our data does not include conference proceedings and books, which play an important role in MCS and SSH, respectively.

Figure 5 shows for each of our five disciplines the CDFs of the level, depth, and dependence of the citation impact of highly cited publications. The results are in line with the statistics presented in Table 2. The distribution of the level of citation impact almost coincides for the five disciplines. The values of depth and dependence tend to be substantially lower in MCS and SSH than in BHS, LES, and PSE.

Case study of the field of scientometrics

To provide a more detailed demonstration of our three-dimensional framework for characterizing the citation impact of publications, we now present a case study in which the framework is applied to publications in the field of scientometrics. As explained above, 4,047 clusters of publications were obtained using an algorithmic methodology. One of these clusters can be considered to represent the field of scientometrics. We selected all 182 highly cited publications (i.e., publications with at least 100 citations) in this cluster. For these publications, we calculated the level, depth, and dependence of their citation impact.

Figure 6 presents the depth and dependence of the 182 publications in a scatter plot. The dashed vertical and horizontal lines indicate the median depth (1.71) and median dependence (1.24), respectively. Based on the dashed lines, four quadrants were obtained. In each of these quadrants, we selected a publication for a more detailed analysis. Hence, we selected a publication with a low depth and a low dependence (referred to as P1), a publication with a high depth and a low dependence (P2), a publication with a low depth and a high dependence (P3), and a publication with a high depth and a high dependence (P4). We selected publications with which we are sufficiently familiar ourselves, so that we are able to offer a detailed interpretation of the citation impact of the selected publications. Table 3 lists the four selected publications and reports the level, depth, and dependence of their citation impact and their number of references. In addition, Figure 7 shows for each of the selected publications the distribution of the number of co-citation links and the number of bibliographic coupling links between the selected publication and the publications citing the selected publication.

Table 3. The four selected publications, the level, depth, and dependence of their citation impact, and their number of references.

	P1	*P2*	*P3*	*P4*
Authors	N. J. van Eck & L. Waltman	J. E. Hirsch	L. Egghe	M. Thelwall

Title	Software survey: VOSviewer, a computer program for bibliometric mapping	An index to quantify an individual's scientific research output	The Hirsch index and related impact measures	Extracting macroscopic information from Web links
Journal	Scientometrics	Proceedings of the National Academy of Sciences of the USA	Annual Review of Information Science and Technology	Journal of the American Society for Information Science and Technology
Year	2009	2005	2010	2001
Level	273	2518	116	107
Depth	1.19	5.89	1.34	6.75
Dependence	1.07	0.06	8.26	5.68
Total no. of ref.	37	6	256	65
No. of ref. in data	26	4	175	43

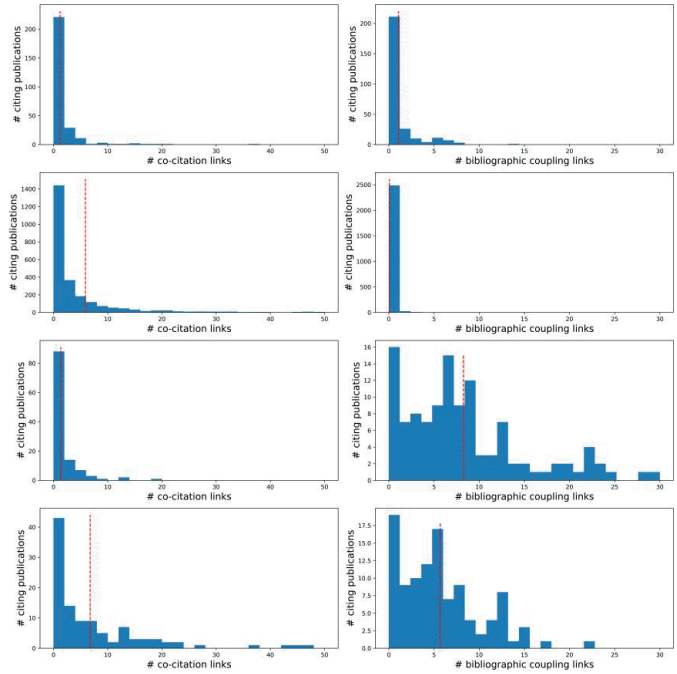


Figure 7. Distribution of the number of co-citation links and the number of bibliographic coupling links between a selected publication and the publications citing the selected publication. The dashed vertical line indicates the mean of the distribution. The mean of the distribution of the number of co-citation links corresponds to the depth of the citation impact of a selected publication. The mean of the distribution of the number of bibliographic coupling links corresponds to the dependence of the citation impact of a selected publication. Subfigures in the first row represents P1, the second row P2, and so forth.

Publication P1 is the article, co-authored by one of us, that introduced the popular VOSviewer software for visualizing bibliometric networks. VOSviewer is used in a large number of publications in many different research fields. Many publications that use VOSviewer cite P1. In our data, P1 has been cited 273 times. Publications that use VOSviewer typically present a bibliometric analysis of the scientific literature in a specific research field or on a specific research topic. Such publications use VOSviewer as a tool for bibliometric visualization. They usually do not aim to develop new bibliometric methods or tools. Consequently, most publications citing P1 do not contribute to the methodological literature on bibliometric visualization. Publications citing P1 therefore tend to refer only sparsely to other publications

on bibliometric visualization. This is reflected by the relatively low depth and dependence of P1. The depth of P1 equals 1.19, which indicates that a publication citing P1 on average has 1.19 co-citation links with P1. In other words, when a publication cites P1, there will on average be 1.19 publications in which the publication is cited together with P1. This means that publications citing P1 are only weakly connected to each other by citation relations. It shows that P1 has a broad but not so deep citation impact. P1 has a dependence of 1.07, indicating that the average number of bibliographic coupling links between each of P1's citing publications and P1 itself is 1.07. Hence, when a publication cites P1, there will on average be 1.07 publications that are cited both by this publication and by P1. Figure 7 shows that there are a limited number of publications citing P1 that have a more substantial number of co-citation links or bibliographic coupling links with P1. Unlike most publications citing P1, these are likely to be publications that contribute to the methodological literature on bibliometric visualization.

Publication P2 is the article published in 2005 by Jorge Hirsch in which he introduced the h-index. This is an extremely influential publication. Indeed, with 2518 citations, P2 is by far the most highly cited publication in the field of scientometrics in our data. There are a large number of publications that present studies of the h-index, propose alternatives to the h-index, or report bibliometric analyses in which the h-index is applied. In the field of scientometrics, one may argue that P2 has been the starting point of a new subfield of research focused on studying bibliometric indicators of the performance of individual researchers (or, alternatively, one could suggest there is an h-bubble; see Rousseau, García-Zorita, & Sanz-Casado, 2013). P2 has a high depth of 5.89. Hence, on average, publications that cite P2 have 5.89 co-citation links with P2. This shows that publications citing P2 are strongly connected to each other by citation relations, which reflects the central position of P2 in a highly active subfield of research. The dependence of P2 is very low. The average number of bibliographic coupling links between publications that cite P2 and P2 itself is only 0.06. In other words, publications citing P2 hardly cite the references of P2. This suggests that P2 does not only have a central position in a specific subfield of research, but that it can be considered a foundational publication in this subfield. However, it is important to be aware that P2 has only a very limited number of references (see Table 3), which means that almost by necessity its dependence will be low. The small number of references of P2 may be seen as additional evidence of the foundational role of this publication. Indeed, according to the author of P2, earlier work did not play 'a major role' in the development of his ideas (Tahamtan & Bornmann, 2018, Appendix A.7). On the other hand, the small number of references of P2 could also be argued to reflect a lack of generosity in the referencing behaviour of the author of P2.

Publication P3, published in 2010, is a review article about the h-index and other related bibliometric indices. P3 has been cited 116 times in our data. It includes 256 references, of which 175 point to publications included in our data. The large number of references reflects the voluminous literature on the h-index published between 2005 and 2010. P3 has a high dependence of 8.26. Hence, when a publication cites P3, it will on average also cite 8.26 references of P3. As can be seen in Figure 7, some publications citing P3 even cite more than 20 of P3's references. The high dependence of P3 indicates that P3 builds on a large body of literature and that the citation impact of P3 is strongly dependent on this literature. This reflects that, as a review article, P3 does not make an original scientific contribution. It is sometimes suggested that researchers have the tendency to cite review articles instead of citing the underlying original works, but the high dependence of P3 shows that this is not the case for P3. The depth of P3 equals 1.34, which is just below the median depth of the 182 highly cited publications in the field of scientometrics. Publications that cite P3 on average have 1.34 co-citation links with P3, indicating that publications citing P3 are only relatively weakly connected to each other by citation relations. This may be due to the gradual decline in the

interest of the scientometric community in the h-index. It also shows that P3 has not developed into a canonical reference for publications dealing with the h-index. This may partly be explained by the fact that around 2010 a number of review articles about the h-index were published more or less at the same time.

Publication P4 is about the extraction of macroscopic information from Web links. This publication deals with a topic in field of webometrics, which partly overlaps with the field of scientometrics. P4 was published in 2001. It has received 107 citations in our data. As can be seen in Figure 6, P4 is a rather unique publication in the scientometric literature, because it combines a high depth (6.75) with a high dependence (5.68). This means that publications citing P4 have lots of citation relations both with each other and with the references of P4. As can be seen in Table 3, the number of references of P4 is not exceptionally large, making P4's high dependence even more noteworthy. The high depth of P4 suggests that P4 makes an important contribution to a relatively narrow but densely connected area of research. On the other hand, the high dependence of P4 seems to indicate that P4 should not be regarded as a pioneering publication. The citation impact of P4 is strongly dependent on earlier publications. Hence, P4 can be considered to make an important incremental contribution but not a highly innovative one.

Discussion and conclusion

We propose a three-dimensional framework for characterizing the citation impact of scientific publications. The proposed framework makes a distinction between the level, depth, and dependence of the citation impact of a publication. The level of citation impact is operationalized by the number of citations a publication has received. The depth of the citation impact of a publication is quantified by calculating the average number of co-citation links between the publication and its citing publications. The more strongly a publication's citing publications are connected to each other by citation relations, the higher the depth of the citation impact of the publication. The dependence of the citation impact of a publication is quantified by calculating the average number of bibliographic coupling links between the publication and its citing publications. The more a publication and its citing publications refer to the same earlier publications, the higher the dependence of the citation impact of the publication.

In a traditional one-dimensional perspective on citation impact, the number of citations received by a publication is used as an indicator of the impact of the publication on later publications. Our three-dimensional framework offers a more detailed understanding of the citation impact of a publication. It enables us to make a distinction between publications that have a deep impact concentrated in one specific field of research and publications that have a broad impact scattered over different research fields. It also allows us to distinguish between publications that are strongly dependent on earlier work and publications that make a more independent scientific contribution.

In the field of scientometrics, we find that the article in which the h-index was introduced has a high depth and a low dependence. This reflects the role of this article as the starting point of a new subfield of research within the field of scientometrics. On the other hand, a review article on the h-index has a high dependence, which shows the strong reliance of this article on earlier works. A high dependence can be expected to be a typical feature of review articles. The article in which the VOSviewer software was introduced has a low depth, reflecting its broad but not so deep impact. Finally, an article in the field of webometrics has a high depth and a high dependence, indicating that this article contributes to a strongly cumulative research area, but that it does not play a pioneering role in this area.

There are various directions for future research. In future work, our citation impact framework could be analyzed in more detail, for instance by studying its mathematical properties and by carrying out additional case studies. Also, the effect of author self-citations in our citation

impact framework could be analyzed. In addition, different ways in which the indicators provided by our citation impact framework are presented to users could be tested and compared. More generally, ideas similar to the ones proposed in this paper could be explored at the level of researchers or research groups rather than at the individual publication level. Finally, the distinction between cumulative research and more independent research could be studied in alternative ways. Research areas that are of a strongly cumulative nature for instance could be identified by searching for densely connected subnetworks in a citation network of publications.

Acknowledgments

Yi Bu would like to thank Ying Ding, Xianlei Dong, and Jian Xu for their comments on an earlier draft of this paper. The authors are grateful to Lutz Bornmann and to three anonymous reviewers for their comments and suggestions.

References

Chen, P., Xie, H., Maslov, S., & Redner, S. (2007). Finding scientific gems with Google's PageRank algorithm. *Journal of Informetrics, 1*(1), 8–15.

Clough, J. R., Gollings, J., Loach, T. V., & Evans, T.S. (2015). Transitive reduction of citation networks. *Journal of Complex Networks, 3*(2), 189-203.

Ding, Y., Liu, X., Guo, C., & Cronin, B. (2013). The distribution of references across texts: Some implications for citation analysis. *Journal of Informetrics, 7*(3), 583-592.

Funk, R. J., & Owen-Smith, J. (2016). A dynamic network measure of technological change. *Management Science, 63*(3), 791-817.

Huang, Y., Bu, Y., Ding, Y., & Lu, W. (2018a). Number versus structure: Towards citing cascades. *Scientometrics, 117*(3), 2177-2193.

Huang, Y., Bu, Y., Ding, Y., & Lu, W. (2018b). *Direct citations between citing publications*. arXiv:1811.01120.

Rousseau, R., García-Zorita, C., & Sanz-Casado, E. (2013). The h-bubble. *Journal of Informetrics, 7*(2), 294-300.

Tahamtan, I., & Bornmann, L. (2018). Creativity in science and the link to cited references: Is the creative potential of papers reflected in their cited references? *Journal of Informetrics*, *12*(3), 906-930.

Waltman, L. (2016). A review of the literature on citation impact indicators. *Journal of Informetrics, 10*(2), 365-391.

Waltman, L., & Van Eck, N. J. (2012). A new methodology for constructing a publication-level classification system of science. *Journal of the American Society for Information Science and Technology, 63*(12), 2378-2392.

Waltman, L., & Van Eck, N.J. (2018). *Field normalization of scientometric indicators*. arXiv:1801.09985.

Waltman, L., & Yan, E. (2014). PageRank-related methods for analyzing citation networks. In Y. Ding, R. Rousseau, & D. Wolfram (Eds.), *Measuring scholarly impact: Methods and practice* (pp. 83-100). Springer.

Wan, X., & Liu, F. (2014). WL-index: Leveraging citation mention number to quantify an individual's scientific impact. *Journal of the Association for Information Science and Technology, 65*(12), 2509-2517.

Wu, L., Wang, D., & Evans, J. A. (2019). Large teams develop and small teams disrupt science and technology. *Nature, 566*(7744), 378.

Zhu, X., Turney, P., Lemire, D., & Vellino, A. (2015). Measuring academic influence: Not all citations are equal. *Journal of the Association for Information Science and Technology, 66*(2), 408-427.

Higher Education's Role in Chinese National Innovation System: A Perspective of University-Industry Linkages

Yu Chen[1], Jiawei Han[2], Zhaohui Xuan[3] and Wen Gao[4]

[1] *cheny@casted.org.cn* [2] *hanjw@casted.org.cn* [3] *xuanzh@casted.org.cn*
Chinese Academy of Science and Technology for Development, No.8 Yuyuantan South Road, Beijing, 100038 (China)

[4] *gaow@tsts.org.cn*
Tianjin Science and Technology Statistics and Development Research Center, 305 Nanjing Road, Tianjin, 300052 (China)

Abstract
Higher Education Innovation Survey, consisting of University Innovation Survey and Related Science and Technology Personnel Survey, has been conducted by the Chinese government since 2016. Using the newly completed survey data, this paper focuses on Chinese universities' performance in national innovation system and analyses the linkages between universities and enterprises in terms of personnel training, university-industry collaboration and technology transfer. We find out that firstly, Chinese universities have made great achievements in the cultivation of students' innovative and entrepreneurial abilities by jointly training students with enterprises, establishing internship training bases and encouraging students to participate in research projects. Secondly, universities have developed a close relationship with enterprises by cooperating in scientific research projects and jointly establishing scientific research institutes. Thirdly, universities have tried to facilitate technology transfer by establishing technology transfer agencies and websites and have become an important technology source to enterprises. Finally, there are still some disputes and obstacles in the cooperation between universities and enterprises, especially during the process of technology transfer.

Introduction

As economies become increasingly knowledge-based, scientific and technological efforts become essential determinants of industrial performance and international competitiveness (OECD, 1999 & 2001). As a primary sector of national knowledge production and technology research, the Higher Education Institutions (HEIs) are increasingly connected with economic and social development and have become an important part of the national innovation system. Therefore, the analysis of HEIs' innovation activities and their efficiency has become an important and popular topic of the national innovation system research around the world.

In order to reflect the strength and effectiveness of universities in supporting and promoting enterprise innovation, China's Ministry of Education and Ministry of Science and Technology have launched the *Higher Education Innovation Survey* since 2016 to assess the flow of innovation elements and linkages between universities and enterprises and provide a

statistical basis for formulating innovation policies. Based on the theoretical framework of the national innovation system, this paper uses the newly completed *Higher Education Innovation Survey* data to comprehensively analyse the performance of Chinese universities in personnel training, university-industry collaboration, and technology transfer.

Data and Methodology

The cross-section data used in this paper was obtained from the second national *Higher Education Innovation Survey* jointly conducted by the Ministry of Education and the Ministry of Science and Technology of China in 2017. The survey has been carried out annually since 2016 and can be divided into two levels: University Innovation Survey and Related Science and Technology Personnel Survey. The University Innovation Survey (UIS), a full sample survey finished by the research management office of each university, focuses on universities' role in supporting and promoting enterprise innovation through the channels of talent training, university-industry collaboration and technology transfer. The Related Science and Technology Personnel Survey, which is the sampling survey of university teachers (TS) and graduate students (GSS), pays attention to individual behaviours of research and innovation, their perceptions of different kinds of innovation activities, and their understandings of the obstacles and policy appeals during the cooperation with enterprises.

Table 1. Measurement of universities' functions in the national innovation system

Dimensions	*Indicators*	*Sources*
I. Talent training	Number of graduate students	UIS
	Joint education	TS
	Students participating in research projects	TS; GSS
	Internship training bases	UIS
	Entrepreneurship projects and courses	UIS; GSS
	Innovation and entrepreneurship awards	UIS
II. University-industry collaboration in R&D	Cooperation in research projects	UIS
	Joint research institutes	UIS
	Disputes during co-research	TS
III. Technology transfer	Agencies and websites	UIS
	Technology transfer contracts	UIS
	Patent transfer	UIS
	Obstacles to technology transfer	TS

We use the data of 2016 both at the university level and the individual level to analyse the characteristics of innovation activities in Chinese universities focusing on university-industry linkages. 377 HEIs at the university level were selected as the analytical sample, covering all different types of universities and with all varieties of disciplines including comprehensive, science and engineering, agriculture, forestry, and medicine, etc. At the individual level, we use the data of full-time teachers with various professional titles, who undertook both

teaching and research tasks in 2016, and postgraduates including masters and PhDs in various disciplines. 11,690 teachers and 2,010 students are included in our sample.

A series of indicators, both quantitative and qualitative, are selected in this paper to describe the performance of universities and their linkages with enterprises. Table 1 presents all the indicators in our analysis which can be classified into three categories: talent training, university-industry collaboration in R&D and technology transfer. In the following sections, we employ cross-tab analysis combined with frequency distribution analysis and show our findings through statistical analysis charts.

Main findings

Talent Training

Talent training is the core function of higher education in society. After the rapid development over the past few decades, China has established the world's largest higher education system (MoE & MOST, 2018). Chinese universities have cultivated a large number of talents with knowledge and skills for business and social development. In 2016, 1,665,441 students graduated from 377 universities with bachelor's degree, 433,088 students with master's degree and 45,349 students with doctoral degree, of which 54.5%, 52.6% and 20.5% chose to work in enterprises.

In order to train more talents suitable for the development needs of enterprises, Chinese universities have increasingly attached importance to the joint training of students with enterprises and social organizations. For the question "Have you ever cooperated with enterprises, government departments or institutions in jointly training students in the past three years?", more than 20% of the teachers chose "frequently" or "always", and "occasionally" constituted 35.5%. Most teachers believed that joint training of students with enterprises had achieved good results, and the proportion of choosing "good effect" and "very good effect" reached 73% combined (Table 2).

Table 2. Frequency and effect of joint education Unit: %

	Effect of joint education				
Frequency	*No effect*	*Small effect*	*Good effect*	*Very good effect*	*Total*
Never	5.2	11.1	25.6	2.1	44.1
Occasionally	0.4	8.2	24.7	2.1	35.5
Frequently	0.1	1.5	12.3	2.8	16.6
Always	0.0	0.2	1.5	2.0	3.8
Total	5.7	21.1	64.1	9.1	100 (N=11690)

By encouraging students to participate in research projects from enterprises, universities could better develop students' market-oriented R&D capabilities. Among the graduate students surveyed, 87% of the students had participated in research projects. Specifically, 51% of the students participated in 1 or 2 projects, 17% in 3 projects, and 19% in 4 or more projects (Figure 1). About one third of graduate students had been involved in research projects commissioned by enterprises or in cooperation with enterprises. Among these projects, about

32% were mainly aimed at solving practical production problems, 20% for product development and 19% for frontier exploratory theoretical research (Figure 2).

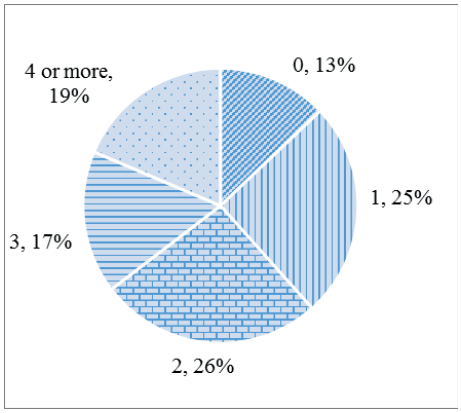


Figure 1. Distribution of graduate students, by number of projects

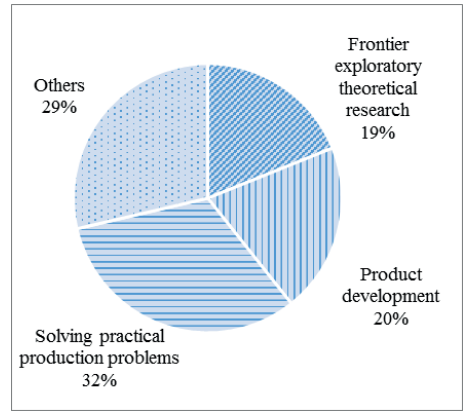


Figure 2. Types of corporate research projects that graduate students participated in

Encouraged by the Chinese government's "mass entrepreneurship and innovation" policy, Chinese universities have attached great emphasis to the cultivation of students' innovative and entrepreneurial abilities. By the end of 2016, Chinese universities had established numerous on-campus and off-campus internship training bases. In general, each university had an average of 25.5 on-campus internship training bases and 199.4 off-campus internship training bases (Figure 3). The average numbers of bases in different types of universities were almost the same, except for medical universities. The number of on-campus and off-campus internship training bases in medical universities was about half of the numbers in other types of universities.

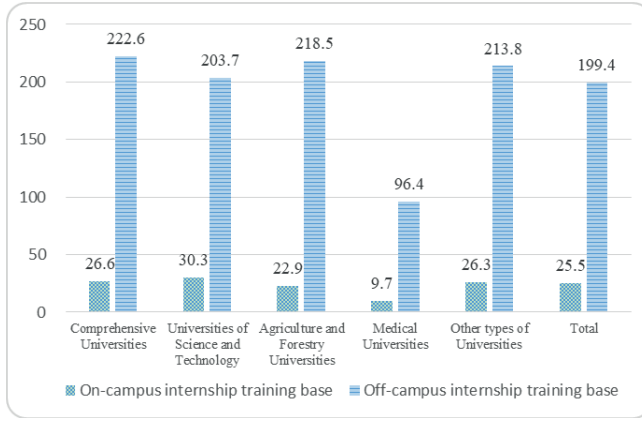


Figure 3. Number of internship training bases per university

On average, each university had 83.8 entrepreneurship projects and 30 entrepreneurship courses in 2016. Overall, 34.5% of graduate students had participated in innovative and entrepreneurial courses, 21.9% had participated in the innovation and entrepreneurship associations, and 20.2% had tried to start their own businesses (Table 3). Government and universities' support for students' innovative and entrepreneurial activities had achieved remarkable results. In 2016，Students won a total of 19,787 National Innovation and Entrepreneurship Awards and 36,940 provincial innovation and entrepreneurship awards respectively.

Table 3. Percentage of students participating in innovative and entrepreneurial activities

Questions	*Yes (%)*	*No (%)*
Have you ever taken any innovative and entrepreneurial course?	34.5	65.5
Have you ever participated in any innovation and entrepreneurship association?	21.9	78.0
Have you ever tried to start your own business?	20.2	79.8

University-Industry Collaboration in R&D

University-industry collaboration in R&D, or briefly co-research, offers an important way for universities to support the development of the enterprise sector by combining the technological supply with the market demand. In China, business funds have always been a crucial source for internal R&D expenditure of higher education, representing 29.0% in 2016, which was much higher than that in most developed countries, such as 2.7% in Japan, 5.5% in the U.S. and 15.0% in Germany[i]. This ensures a continuous and strong collaboration between universities and industry.

Universities have developed a close relationship with enterprises by cooperating in scientific and technological research projects. In 2016, Chinese universities made a great number of project contracts with enterprises, the value of which reached 36.9 billion Yuan. Specifically, just as Figure 4 shows, the value of contracts with enterprises in the same province accounted for the largest share of the total, which was 49%, slightly higher than that of the inter-provincial contracts (46.70%), and much higher than that of the overseas contracts (4.30%). The difference was even more obvious when we look at the number of enterprises that cooperated with universities. On average, each university undertook projects with 133 enterprises in the same province, 88 enterprises in other provinces and 2 overseas enterprises (Figure 5). This might provide possible evidence for the theory of the role of geographical proximity in universities' spillovers, or in this context, co-research (Fritsch, 2010). With the increase of the distance, the efficiency of the communication and the benefit from collaboration decrease (Branstetter, 2001; Boschma, 2005). Besides, universities of science and technology and comprehensive universities cooperated with more enterprises (348.5 and 336.5 on average respectively) than other types of universities (less than 200).

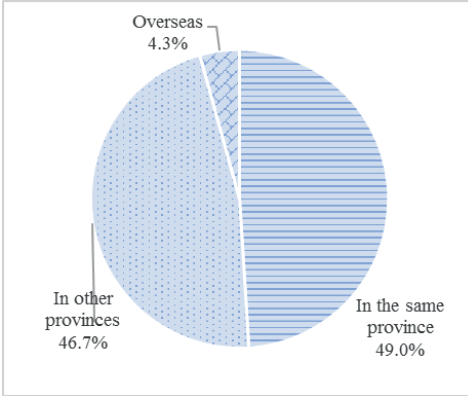


Figure 4. Shares of contract value, by location of enterprises

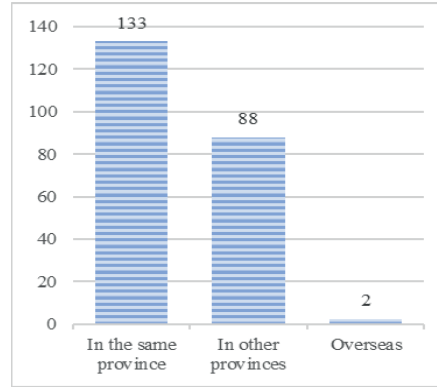


Figure 5. Average number of enterprise partners of each university, by location of enterprises

Scientific research institutes jointly built by universities and enterprises are good carriers for co-research. By the end of 2016, 261 universities, or about 70% of the total, had established joint scientific research institutes with enterprises, the number of which reached 2241 in total. There were 110 universities standing out of these universities, each owning at least six joint research institutes, implying a wide and strong links between universities and industries (Figure 6). The number of joint research institutes varied across universities with different levels and types. In general, universities that are recognized as "211-project" universities[ii] were more likely to develop close relationships with enterprises due to their better capacity, reputation and network. The percentage of "211-project" universities having more than one institute (70%) was significantly higher than that of "non-211-project" universities (50.5%) (Figure 7). The type of universities also played a role. Universities of which major research areas are science and engineering, including universities of science and technology, comprehensive universities and agricultural and forestry universities, tended to have more need for linkages with industries and build more research institutes with enterprises, with the average number being 8.8, 7.8 and 5.8 respectively. One exception is medical universities, with the average number being only a bit higher than other types of universities[iii] (Figure 8). The co-research brought about lots of scientific and technological achievements. In terms of invention patents, universities' teachers, as the first inventor, filed 116,461 invention patent applications and got 53,027 patents granted, of which 23.3% and 27.3% were from the projects jointly undertaken by universities and enterprises in 2016.

However, there were some disputes that universities' teachers would confront during the process of co-research, such as not getting funds in time, disputes in intellectual property rights and income distribution, and disagreements on delivery schedule, ideas, criteria for completion and management of funds. But only a few teachers had been faced with these problems. Less than one fourth of teachers reported that these problems occurred often or frequently (Figure 9).

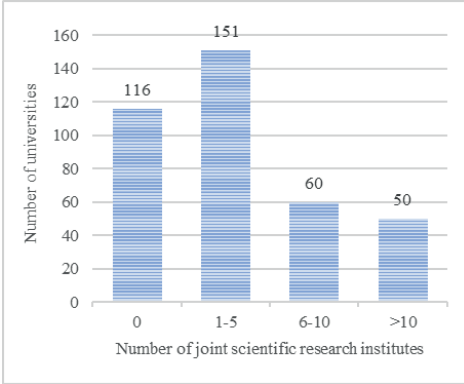


Figure 6. Number of joint scientific research institutes

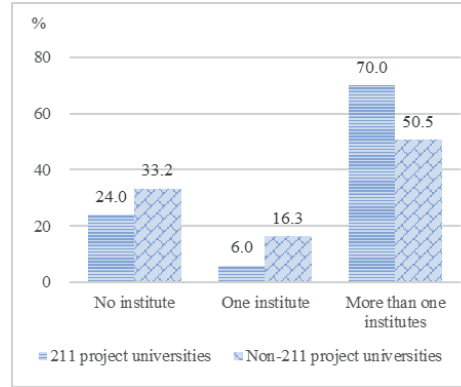


Figure 7. Percentage of universities with joint scientific research institutes

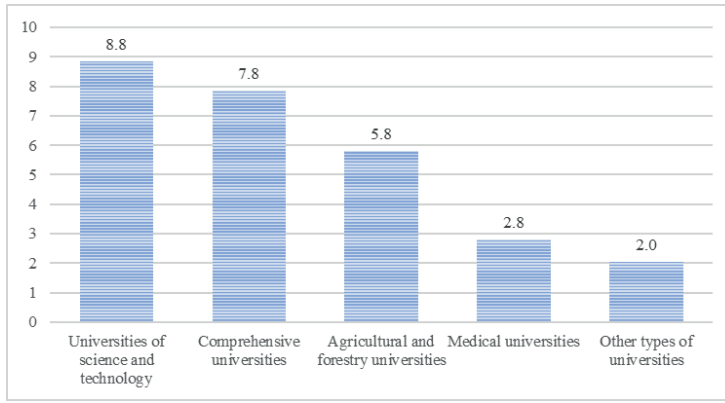


Figure 8. Average number of joint scientific research institutes, by type of universities

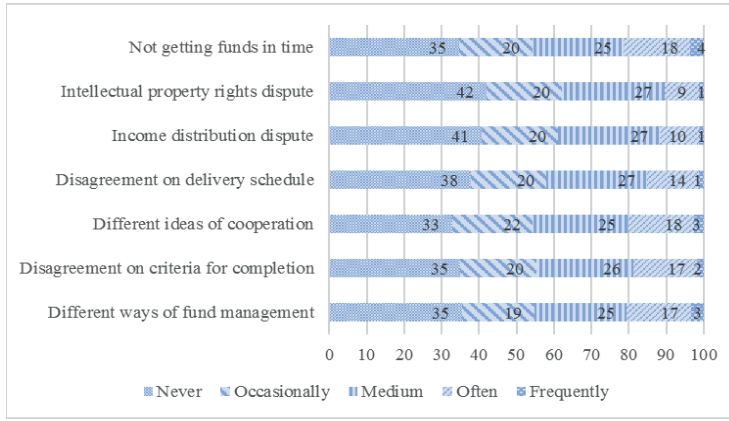


Figure 9. Disputes during co-research

Technology Transfer

Apart from co-research, direct technology transfer, which is the flow of technology from universities to enterprises, is also an effective way to play the advantage of universities' research. Numerous universities have made great efforts to facilitate technology transfer by establishing technology transfer agencies and websites. By the end of 2016, 264 universities, or around 70% of the total, had set up specialized agencies and 190 universities, or about 50%, had created relevant websites. Furthermore, universities with stronger R&D capabilities and with focus on science and engineering are more likely to have technology transfer agencies and websites, perhaps due to the large demand for professional services and more resources to mobilize. Specifically, the share of "211-project" universities having specialized agencies (82%) was 16.3 percentage points higher than that of "non-211-project" universities (65.7%), and the gap was even larger with regard to websites (20 percentage points). Universities, of which major research areas are science and engineering, are more likely to have technology transfer agencies and websites except that medical universities tended to have comparatively less websites (Figure 10). Apart from this, there might be a lack of professional staff for technology transfer services. More than 70% of the specialized agencies had less than eight staff and lots of them had only one or two staff, which apparently were incapable of providing complete, systematic and comprehensive services for technology transfer.

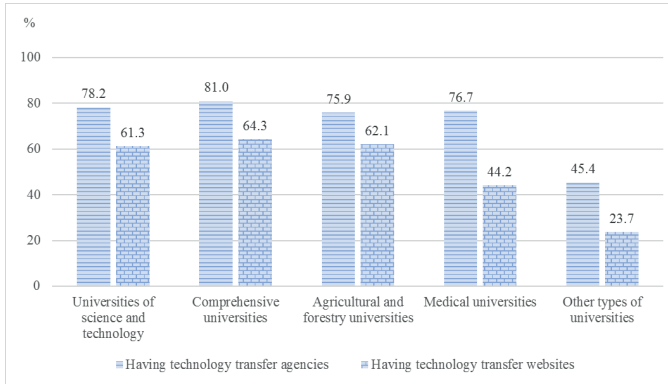


Figure 10. Percentage of universities having technology transfer agencies and websites, by type of universities

Universities have become important technology suppliers to enterprises. In 2016, universities signed a total of 7,755 technology transfer contracts valued at 4.35 billion Yuan. There were two characteristics about the contracts. Firstly, the number of contracts declined by the order of transaction partners, i.e. enterprises in the same province, in other provinces and overseas. Intra-provincial contracts constituted 61.7% of the total, inter-provincial contracts 37.1% and overseas contracts 1.1%. Further, the difference between the proportions of intra-provincial and inter-provincial contracts was smaller for "211-project" universities, about 10% difference, versus 35% difference for "non-211-project" universities. Secondly, on average, "211-project" universities signed more contracts, 31.5 per university, than "non-211-project" universities, 16.7 per university (Table 4). Together with the first feature, this might indicate a

larger radiation effect of "211-project" universities' scientific and technological research, especially to enterprises in other provinces.

Table 4. Universities' technology transfer, by type of contracts

Types of contracts	*Total*		*"211-project" universities*		*"non-211-project" universities*	
	No. of contracts	*Percentage (%)*	*No. of contracts*	*Average*	*No. of contracts*	*Average*
Intra-provincial contracts	4788	61.7	1689	17.1	3099	11.2
Inter-provincial contracts	2878	37.1	1384	14.0	1494	5.4
Overseas contracts	89	1.1	46	0.5	43	0.2
Total	7755	20.6	3119	31.5	4636	16.7

Patent transfer is an important form of technology transfer in universities. In 2016, Chinese universities implemented up to 11,425 technologies of patents in a variety of ways, and universities of science and technology contributed to the largest share followed by comprehensive universities and agricultural and forestry universities. In general, implementing technologies independently was the most popular way among universities, representing 28.9% of the total, followed by patent licensing and transfer of patent rights at 28.0% and 24.1% respectively. Moreover, different types of universities showed different preferences. The most popular way of medical universities is to implement technologies independently. Universities of science and technology showed similar preferences over independent implementation and patent licensing. Agricultural and forestry universities used patent licensing most, and transfer of patent rights is more commonly used by comprehensive universities (Table 5).

Table 5. Percentage of ways of patent transfer, by type of universities

	Ways of patent transfer					
Types of universities	*Independent Implementation (%)*	*Transfer of patent rights (%)*	*Patent Licensing (%)*	*Converting technology value into enterprise shares (%)*	*Joint Implementation (%)*	*Total No.*
Universities of science and technology	31.5	18.3	28.5	6.8	14.9	6772
Agricultural and forestry universities	26.9	15.5	40.4	7.8	9.5	1687
Medical universities	69.2	14.7	9.4	1.0	5.7	299
Comprehensive universities	12.4	48.7	23.0	4.8	11.1	1919
Other types	35.6	37.3	15.4	2.8	9.0	748
Total	28.9	24.1	28.0	6.2	12.8	11425

Despite that, there still existed some obstacles hindering universities' technology transfer concerning information, incentives, financial support, policies, market potential and intermediaries. The lack of financial and policy support were the two most serious problems, and more than 60% of the teachers agreed that these two problems had blocked technology transfer to a great extent. Relatively, the lack of market potential and well-organized intermediary institutions were less serious, and less than 50% of the teachers considered them as big problems (Figure 11). Combined with the analysis of the disputes during co-research, problems in technology transfer were much worse and needed more attention.

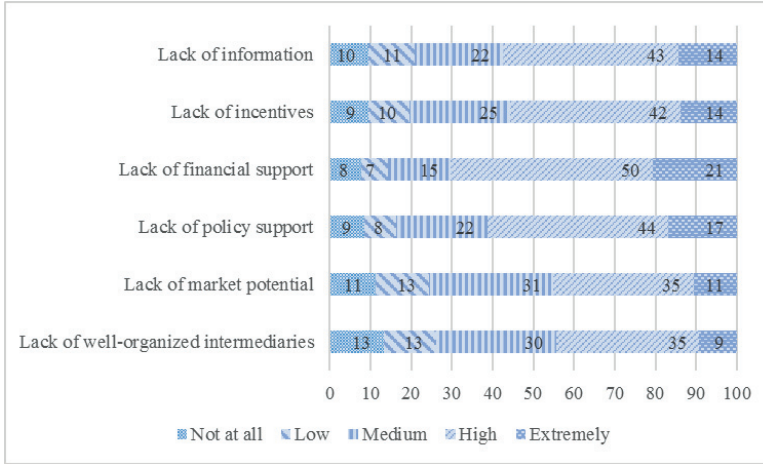


Figure 11. Obstacles to technology transfer

Conclusions

Due to the indispensable role of universities in promoting enterprise innovation, Chinese universities are trying to strengthen their linkages with enterprises and stimulate economic vitality through talent training, co-research and technology transfer. Using the data of 2016 from the national *Higher Education Innovation Survey*, we have presented the status quo of Chinese universities' linkages with enterprises in terms of innovation and drawn some conclusions.

First of all, Chinese universities have made great achievements in the cultivation of students' innovative and entrepreneurial abilities by jointly training students with enterprises, establishing internship training bases and encouraging students to participate in research projects. By the end of 2016, each university had an average of 25.5 on-campus internship training bases and 199.4 off-campus internship training bases. And each university had 83.8 entrepreneurship projects and 30 entrepreneurship courses on average. Among graduate students surveyed, there were 87% and 34.5% had participated in research projects and innovative and entrepreneurial courses, respectively.

Secondly, universities have developed a close relationship with enterprises by cooperating in scientific research projects and jointly establishing scientific research institutes. By the end of

2016, about 70% of the universities had established joint scientific research institutes with enterprises. And on average, each university undertook projects with 133 enterprises in the same province, 88 enterprises in other provinces and 2 overseas enterprises.

Thirdly, universities have tried to facilitate technology transfer by establishing technology transfer agencies and websites and have become an important technology source to enterprises. By the end of 2016, around 70% of universities had set up specialized agencies and about 50% had created relevant websites. Universities signed a total of 7,755 technology transfer contracts valued at 4.35 billion Yuan and implemented up to 11,425 technologies of patents in a variety of ways in 2016.

Finally, we observe some disputes and barriers in the cooperation between universities and enterprises, especially during the process of technology transfer. More than 60% of the teachers agreed that the lack of financial and policy support were the two most serious problems for technology transfer. In order to fully and effectively exploit the market value of universities' technologies, relevant S&T services development should be accelerated, such as technology finance, patent assessment and incentive policies.

References

Boschma, R. (2005). *Proximity and innovation: a critical assessment.* Regional studies, 39(1), 61-74.

Branstetter, L. G. (2001). *Are knowledge spillovers international or intranational in scope? Microeconometric evidence from the US and Japan.* Journal of International Economics, 53(1), 53-79.

Fritsch, M. (2001). *Co-operation in regional innovation systems.* Regional Studies, 35(4), 297-307.

Ministry of Education & Ministry of Science and Technology P. R. C. (2018). *China's higher education institutions innovation capability monitoring report 2018*. Beijing: Scientific and Technical Documentation Press.

OECD. (1999). *Science, technology and industry scoreboard 1999: Benchmarking knowledge-based economies.* Paris.

OECD. (2001). *Science, technology and industry scoreboard: Towards a knowledge-based economy.* Paris.

---

[i] Data sources: Calculated from the database of OECD, https://stats.oecd.org/.

[ii] The "211 Project", initiated by the Chinese government in 1995, is aimed at supporting the construction of about 100 key higher education institutions and a number of key disciplines facing the 21st century.

[iii] Other types of universities are universities of which major research areas are art and humanities, including normal universities, universities of political science and law, universities of finance and economics and language universities.

Impact Indicator on Measuring Multi-Dimension Technological Convergence

Bowen SONG[1]，Chunjuan LUAN[1, 2]

[1] bowensong333@163.com

Institute of Humanities & Social Sciences, Dalian University of Technology, Dalian, 116085, China

[2] julielcj@163.com; CLuan@dlut.edu.cn

School of Intellectual Property, Dalian University of Technology, Panjin, 124221, China

Abstract

In the past, more attention has been paid to the characteristics of a single technology field while neglecting the role of domains in converging technology. Indicators of converging technology promise valuable technical fields to those determining R&D priorities. We present an implemented impact index to identify inter-domain relations in Multi-dimension technical fields. We propose convergence impact score (CIScore) to evaluate the ability of convergence in the technical fields. At the same time, the feasibility and practicability of the method are verified by taking nanotechnology as an example. Results in this paper can help us better understand the structure of converging technology, and it is easy for us to identify which technical fields have more convergence potential.

Introduction

With the rapid development and dissemination of “emergence”, many of the accelerated technological changes are regarded as the result of the convergence of previously separate disciplines and domains. By reading the relevant information, it can be found that emergence is not a brand-new notion, it had been mentioned at least for two decades(Song, Elvers, & Leker, 2017; Tunzelmann, 1998). Especially, after the remarkable success of Nanotechnology, Mobile Digital Technology and Biotechnology, Convergence receives great attention. Most governments and technology institutions have undergone or are undergoing major reforms in their organizations in order to respond to the great changes in the era of convergence. The US Department of Commerce, the National Science Foundation and the National Science and Technology Commission jointly proposed the issue about “Converging Technologies for Improving Human Performance”(Roco & Bainbridge, 2003). The EU has also put forward a development plan on " Converging Technologies –Shaping the Future of European Societies "(Nordmann, Bruland, Bibel, & Others, 2004). China's scientific research

institutions have repeatedly emphasized the important role of convergence in technological innovation. In so doing, there is a need to identify the research domains with great convergence potential in the development of technology.

Until now, A systematic approach to evaluate the convergence ability of technical field in specific research topics have not been found. At the same time, it is rare to study the inter-domain relationship in Multi-Dimension converging technology. This paper presents a systematic framework for building a model to identify the convergence relationships between technical fields and designing a “Convergence Impact score” to evaluate the convergence ability of technical field.

It is of great significance to study the convergent phenomena of Multi-Dimension converging technology both for theory and practice. Theoretically, this study is helpful in understanding the formation process of convergence phenomenon and the current technological paradigm of science and technology. Practically, it can provide support for the national scientific and technological layout and provide reasonable R&D orientation for enterprises.

Literature review

Converging Technology

We start with a brief, broad consideration of “Converging Technology” from different perspectives, and then evaluate reasonable measuring standards. On the concept of "Converging Technology", many scholarly fields consider the notion of emergence in various contexts. Since the 1980s, the idea of convergence has received great attention(Bröring, Cloutier, & Leker, 2010). At that time, converging technology sometimes is considered to be the intersection of scientific theory or technology, sometimes to the unification of different needs, sometimes to a common goal that is approached from different directions. Since 2001, the concept of converging technology has gradually been unified due to the convening of the conference which named “Converging Technologies for Improving Human Performance – Nanotechnology, Biotechnology, Information Technology and Cognitive Science”(Roco & Bainbridge, 2003). The term “converging technologies” has taken on a new, specific meaning through the combine of “nanotechnology, biotechnology, information Technology and cognitive Science” as well as the subsequent formulation of “NBIC convergence.”(Nordmann et al., 2004).

“Converging technology” can be associated with cross cooperation between different technical fields, tracing back to Pavitt (Pavitt, 1984), to differentiate form the perspectives of tool sharing. “Converging technology” range from the mixture of knowledge, tools and methods to the concept of discrete items moving toward unity or uniformity or the merging of distinct technologies, devices, or industries into a unified whole(Jeong & Lee, 2015). But the concepts of convergence and fusion are still easily confused. Converging technology is a new field of technology in two different fields to achieve a certain goal or meet a certain demand. However, technology fusion is a new way of combination within the field(Elliott, 2001; Kodama, 1992). Fig 1 shows the

common forms of converging technologies. Borés used statistical methods to study that demand is the main factor of convergence (Borés, Saurina, & Torres, 2003). Various features of converging technology have been discussed as below:

Combination of technologies from two or more different application areas
Belongs to the new technical field
Inter-domain cooperation shows high growth potential
To meet certain industrial needs

In summary, converging technology is the source of innovation and the driving force of development(You, Kim, & Jeong, 2014). Converging research streams is also a necessary attribute for the development of emerging technologies(Rotolo, Hicks, & Martin, 2015). Nowadays, many researches are mining converging technologies from all angles. For instance, Kim M et al. used entropy and gravity concepts as new methodological indexes to investigate technological convergence(Cho & Kim, 2014). Network analysis based on patent data is used to identify the occurrence of technology convergence in terms of its technological domains(Jeong, Kim, & Choi, 2015). However, few studies focus on researched the relations between domains in multi-dimension converging technology, at the same time, i.e., there is a lack of indicators to evaluate the convergence influence in the field of technology.

Criteria for Multi-dimension Convergence impact assessment

The Convergence impact assessment described in this paper is a description of convergence potential in the technical fields. Through this assessment, we can evaluate the possibility of technological convergence in the field. It is necessary to construct a reasonable evaluation index which should have both network and point attribute to describe the technology convergence ability of the field. In the assessment index, network attribute shows the role of domain in the paradigm of target technology. Point attribute reflect domain capabilities and characteristics.

These problems should be considered in the network attributes of evaluating convergence influence. First of all, the convergence indicators need to be able to express the multivariate relationships among domains. the generation of converging technologies must be the intersection of technologies between two or more domains. A number of researchers have introduced method to describe the relationship between domains of converging technologies. But most of them study convergence technology in pairs, ignoring the relationships between technical fields of Multi-dimension converging technologies (Figure 1). For one, we take a quantitative approach using nanotechnology patent data, mining 12 association combinations by Apriori association algorithm. It includes four multi-dimension combinations and eight double-dimension combinations. Readers should recognize that there may be disadvantages to analysis of multivariate relations in converging technologies by using double-dimension method. At the same time, the impact of relationships between technical fields must be quantified accurately.

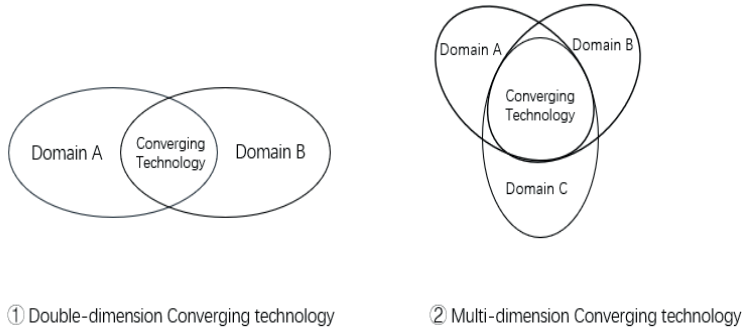


Figure 1. Ways to generate converging technology

When evaluating point attribute in the field, we should consider that convergence impact of technical fields should be hierarchical. This means that the values of convergence with different fields are different. For instance, A is a technical field with high convergence impact index which means it generates converging technologies with many other fields. If convergence occurs between B and A, then B will have the opportunity to generate more converging technologies.

These two attributes are very important in the construction of assessment indicators of convergence impact. Using only a single research paradigm, neither graph theory nor social network analysis can accurately represent the two attributes required by the indicator. We found a compound evaluation model, the improved cross-impact analysis method is used to describe the relationship between domains, eigenvector is introduced to access the point attribute in the technical fields. This is precisely how to solve the problem of convergence by means of convergence.

Data and Methods

Data

This paper aims at exploring convergence between different technical fields, focusing on the multi-dimension convergence relations. Because patent is an important tool for reflecting technological development. Patents in this study are retrieved from the database of Derwent innovations index. The DC code called Derwent classification code is used as the classification basis of technical fields for convergence influence analysis in this study. One of the reasons is that we believe the Derwent classification code employed here is likely to present the technical fields more accuracy then other strategies. Another reason is that DC code is a kind of patent-indexed code designed by Derwent's professionals, and DC-based searching has been employed in some studies(Dirnberger, 2011; Geum, Kim, Lee, & Kim, 2012; Xu, Wang, Liu, & Luan, 2013). There is possibly a third reason: Compared with IPC classification number, DC code can describe each technical field more clearly. The Derwent classification code includes 3 categories and 20 disciplines, it covers all the existing technologies.

Compared with IPC code, DC code is more inclined to be applied and its explanation is clearer. In order to more accurately describe the relationship between technical fields, DC code in Derwent innovations index database is used as the source of domain classification information.

Methods

This paper reflects a patent measurement to identify converging technologies and evaluate convergence impact. Our systematic approach for building a convergence impact model based on multi-dimension converging technology is constructed in four stages (Figure 2).

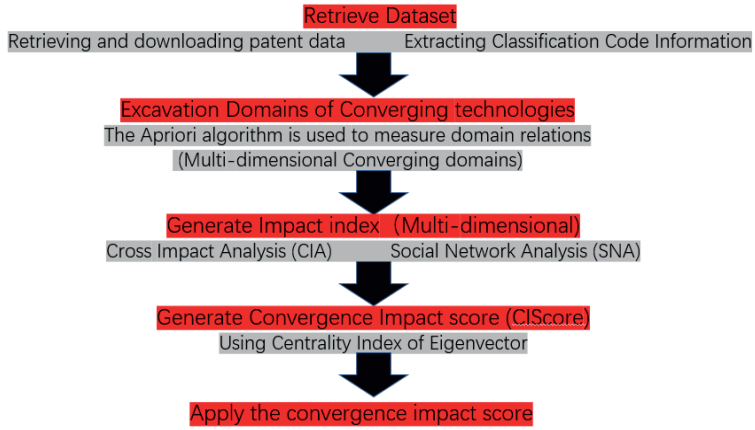


Figure 2. Technology Roadmap

Data retrieval

The first stage is to retrieve the patents from Derwent innovations index database. As per Figure 2, step 1 has been introduced under "Retrieve Dataset". This is the first but most important step. Because different retrieval methods will lead to completely different analysis results.

Take nanotechnology as an example, nanotechnology is one of the most concerned areas in the research of converging technology. The content of nanotechnology is very broad, which can be divided into four parts: metrology and Nano-processes; Nanostructure chemistry and materials; Nanodevices and Nanoelectronics; Nano-medicine and Nano-biotechnology(Porter, Youtie, Shapira, & Schoeneck, 2008). In order to build a retrieval model on such topics, we must fully consider the comprehensiveness of content coverage. In order to ensure the authoritativeness of the analysis results, the latest OECD (The Organisation for Economic Co-operation and Development) retrieval method is selected in this study. For instance, we use the retrieval method of nanotechnology published by OECD (IP= (B82B-001/00 OR B82B-003/00 OR B82Y-099/00 OR B82Y-040/00 OR B82Y-035/00 OR B82Y-030/00 OR B82Y-025/00 OR B82Y-020/00 OR B82Y-015/00 OR B82Y-010/00 OR B82Y-005/00)). It effectively reduces the error caused by using only the simple search term

of "Nano*".

After downloading and cleaning the data, the information of Derwent Classification code in patent data is extracted by program, and then construct the associations between DC into a technology co-occurrence matrix. Technology co-occurrence matrix reports the relationship among different technologies. The matrix construction is the basic work for recognition of convergence network and the construction of convergence impact index.

Excavation Domains of Converging technologies

Existing convergence studies usually directly analyze the whole data set. The analysis method includes statistic occurrence frequency of domains, describing citation network relationships among domains, etc (Choi, Jeong, & Kim, 2015; Jeong et al., 2015; Karvonen & Kässi, 2013). These studies often use global data directly as the analysis objective. Whether or not we should consider the relationship between domains is not entirely Convergence relation. The impact of the linkages between these non-convergent areas on the evaluation results is uncertain, but the impact is inevitable (Brandes, Borgatti, & Freeman, 2016). Therefore, this study extracts the core converging technologies before rendering convergence network.

Apriori association algorithm will be applied to the recognition of converging technologies (Bramer, 2001). We match the patent number (PN) and title information (TI) in the patent information with the transaction ID (TID) in the association algorithm. Each of patents contains a number of different combinations of DC codes. This combination of DC codes is consistent with transactions (TI) in association rules. We can get the converging technology portfolio after setting confidence and support thresholds.

Generate Impact index (Multi-dimension)

The third stage focuses on measuring the relations between technical fields. After mining the convergence relationship in the research objectives, we still need to measure the relationship between domains before drawing the convergence network. There are already some measurements of inter-domain relationships, such as Jaccard coefficient of co-occurrence in evaluation field(Hamers et al., 1989), entropy value method for describing the complexity of inter-domain relations (Cho & Kim, 2014) and other social network indicators(Kranakis, 2013). These indicators have limitations in evaluating multi-dimension convergence relations. On the basis of previous research, we need to propose an impact index of which not only embodies the multi-dimensional relationship between domains, but also describes the influence of domains.

(1) Impact index

As early as 1968, Gordon et al. proposed a method to measure the relationship between them by cross impact analysis (CIA) method(Gordon & Hayward, 1968), later this method was introduced into patent analysis by Geum . (Geum et al., 2012).

In patent analysis, Symbol $N(A \cap B)$ denotes the number of patents jointly owned by technology field A and B while $N(A)$, $N(B)$ represent the total number of patents in domain A and B respectively. Inter-domain impact index $\mathrm{Im}\, pact(A \leftrightarrow B)$ can be calculated by Equation (1) (2).

$$\mathrm{Im}\, pact(A \rightarrow B) = \frac{N(A \cap B)}{N(B)} \tag{1}$$

$$\mathrm{Im}\, pact(B \rightarrow A) = \frac{N(A \cap B)}{N(A)} \tag{2}$$

(2) Limitations of Double-dimension Impact index

Cross impact analysis method (CIA) is a mature measuring method for the evaluation of inter domain relationships in binary convergence networks. If the impact index is low, then the impact is small, and vice versa. This approach describes the differences in interactions between domains.

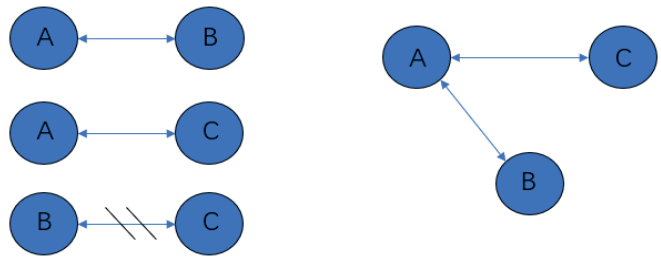


Figure 3. Inter-domain relationships in Multi-dimension Converging technology

However, CIA is not suitable for multi-technology convergence evaluation directly. Because it can cause many problems, such as the decline of accuracy and repeated measurements. For instance, Double-dimension Impact cross impact analysis will repeatedly measure the impact index between domains in a multi-dimension converging technology networks due to its limitations. On the other hand, this evaluation method neglects the existence of Non-direct coupling technology phenomenon in the measurement process (Figure 3). Non-direct coupling technology is a common type of technology in industrial applications. They have to rely on technologies in one field to connect technologies in other fields (Miller et al., 2005; Poudel et al., 2008). The convergence impact needs to be re-measured according to the magnitude of its action intensity, when measuring the magnitude and relationship of multivariate influence among such technological fields.

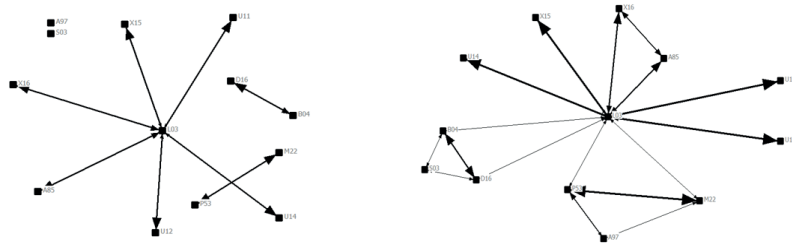
(3) Impact index （Multi-dimension）

In the study of multi-dimension converging technology, full consideration should be given to the role of domains in the Multi-dimension convergence networks. In this study, we introduce the idea of structural hole theory into the evaluation of impact index(Borgatti & Everett, 1997; Burt, 2004). This algorithm will distribute influence according to the proportion of the domain's effort to converging technologies. The calculation of the influence between domains in multi-technology convergence is shown in formula (3).

The above impact index can not only accurately describe the interaction between domains, but also excavate the potential convergence relationship. Formula (3) lists the specific operation process of impact index, influences among domains of different dimensions calculated as $\frac{N(A\cap B\cap ...\cap n)}{N(B)} * \frac{P_A}{P_A+P_C+...+P_n}$ where P_A = the proportion of energy invested in converging technologies in the technical field.

$$
\begin{aligned}
&\mathrm{Im}\, pact_{(multi)}(A\to B)=\sum_{i=2}^{n}\mathrm{Im}\, pact_{(n)}(A\to B)\\
&=\mathrm{Im}\, pact_{(2)}(A\to B)+\mathrm{Im}\, pact_{(3)}(A\to B)+...+\mathrm{Im}\, pact_{(n)}(A\to B)\\
&=\frac{N(A\cap B)}{N(B)}+\frac{N(A\cap B\cap C)}{N(B)}*\frac{P_A}{P_A+P_C}+...+\frac{N(A\cap B\cap ...\cap n)}{N(B)}*\frac{P_A}{P_A+P_C+...+P_n}(n\geq 2) \qquad (3)\\
&P_A=\frac{N(A\cap B\cap C)}{N(A)}\\
&P_C=\frac{N(A\cap B\cap C)}{N(C)}
\end{aligned}
$$

Based on the results calculated by formula, the inter-domain relations of converging technologies in the different dimensions are all significant. We can draw the convergence network according to the obtained impact index, as shown in the figure 4. It is obvious that there are obvious differences between the spectra obtained by Multi-dimension impact index and past methods. More potential relationships are presented in the network.



(a) Double-dimension Inter-domain relations (b) Multi-dimension Inter-domain relations

Figure 4. Inter-domain impact relationship network

Generate Convergence Impact score (CIScore)

In the above, we have successfully demonstrated the relationships between technical fields of the converging technologies, yet failed toreflect the convergence potential or ability on the technical fields. For example, it is still unknown which technical field has the greatest impact on convergence and how much of each technical fields possible convergence. To answer these questions, Convergence Impact score (CIScore) is introduced and used to approach these questions. CIScore index is constructed by combining multiple cross impact analysis and eigenvector centrality. Eigenvector centrality is enabled in the software, Ucinet (Borgatti & Everett, 1992). Eigenvector centrality is an indicator for describing hierarchical attributes of nodes in the network based on the concept that connections to high-scoring nodes contribute more to the score of the node in question than equal connections to low-scoring nodes ("Eigenvector centrality," 2019.1.10; Everett & Borgatti, 1999; Freeman, 1978). Compared with betweenness centrality, point centrality and closeness centrality, eigenvector centrality is more suitable for evaluating the convergence influence in the technical field (Freeman, 1978), main differences between measurement methods present in Figure 5.

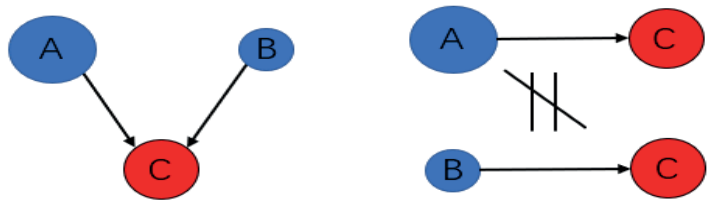


Figure 5. Convergence influence measurement

We measure the network in Figure 4 with eigenvector centrality to get CIScore. Based on the results of eigenvector centrality analysis shown in Table 2, the CIScore clearly react the convergence ability of various fields and also can be used as one of the indicators to evaluate the convergence potential of the fields.

Table 2. Convergence Impact Score (CIScore) of Nanotechnology

No.	Classification Code	CIScore	No.	Classification Code	CIScore
1	L03	0.700	8	B04	0.050
2	U14	0.318	9	D16	0.048
3	U12	0.316	10	P53	0.022
4	U11	0.312	11	M22	0.022
5	X15	0.278	12	S03	0.006
6	A85	0.255	13	A97	0.003
7	X16	0.251			

Conclusions and Discussion

To sum up, this paper offers an impact indicator for measuring Multi-dimension convergence relations, and introduces Convergence Impact Score (CIScore) – providing a quantified metric to describe the convergence ability in descriptive domains. We provide technological process to calculate CIScore and the theory of calculation. The whole calculation process can be divided into four steps: retrieve dataset; excavation domains of converging technologies; generate Impact index（Multi-dimension）and generation of convergence impact score (CIScore).

The presented approach focuses on improving multi-dimension convergence relationships mining and measuring the ability of convergence in the technical field and foresight to judge the appearing of emerging technologies. The intent is that the findings can quickly identify potential relationships between domains and track potential convergence. This result is of great significance for us to make scientific decisions and to evaluate emerging technical fields.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (NSFC) under Grant 71774020/71473028/71603040.

References

Borés, C., Saurina, C., & Torres, R. (2003). Technological convergence: a strategic perspective. *Technovation, 23*(1), 1-13.

Borgatti, S. P., & Everett, M. G. (1992). Notions of Position in Social Network Analysis. *Sociological Methodology, 22*(4), 1-35.

Borgatti, S. P., & Everett, M. G. (1997). Network analysis of 2-mode data. *Social Networks, 19*(3), 243-269.

Bröring, S., Cloutier, L. M., & Leker, J. (2010). The Front End of Innovation in an Era of Industry Convergence: Evidence from Nutraceuticals and Functional Foods. *R & D Management, 36*(5), 487-498.

Bramer. (2001). *Principles of Data Mining*.

Brandes, U., Borgatti, S. P., & Freeman, L. C. (2016). Maintaining the duality of closeness and betweenness centrality ☆. *Social Networks, 44*, 153-159.

Burt, R. S. (2004). Structural holes and good ideas. *American Journal of Sociology, 110*(2), 349-399.

Cho, Y., & Kim, M. (2014). Entropy and Gravity Concepts as New Methodological Indexes to Investigate Technological Convergence: Patent Network-Based Approach. *Plos One, 9*(6), e98009.

Choi, J. Y., Jeong, S., & Kim, K. (2015). A Study on Diffusion Pattern of Technology Convergence&58; Patent Analysis for Korea. *Sustainability, 7*(9), 11546-11569.

Dirnberger, D. (2011). A guide to efficient keyword, sequence and classification search strategies for biopharmaceutical drug-centric patent landscape searches - A human recombinant insulin patent landscape case study. *World Patent Information, 33*(2), 128-143.

Eigenvector centrality. (2019.1.10). from https://en.wikipedia.org/wiki/Centrality#cite_note-ams-11

Elliott, G. (2001). Market-oriented Technology Management—Innovating for Profit in Entrepreneurial Times. *European Journal of Information Systems, 10*(3), 180-180.

Everett, M. G., & Borgatti, S. P. (1999). The centrality of groups and classes. *Journal of Mathematical Sociology, 23*(3), 181-201.

Freeman, L. C. (1978). Centrality in social networks conceptual clarification. *Social Networks, 1*(3), 215-239.

Geum, Y., Kim, C., Lee, S., & Kim, M. S. (2012). Technological Convergence of IT and BT: Evidence from Patent Analysis. *ETRI Journal,34,3(2012-06-01), 34*(3), 439-449.

Gordon, T. J., & Hayward, H. (1968). Initial experiments with the cross impact matrix method of forecasting *. *Futures, 1*(2), 100-116.

Hamers, L., Hemeryck, Y., Herweyers, G., Janssen, M., Keters, H., Rousseau, R., & Vanhoutte, A. (1989). Similarity measures in scientometric research: The Jaccard index versus Salton's cosine formula. *Information Processing & Management An International Journal, 25*(3), 315-318.

Jeong, S., Kim, J. C., & Choi, J. Y. (2015). Technology convergence: What developmental stage are we in? *Scientometrics, 104*(3), 1-31.

Jeong, S., & Lee, S. (2015). What drives technology convergence? Exploring the influence of technological and resource allocation contexts. *Journal of Engineering and Technology Management, 36*, 78-96. doi: 10.1016/j.jengtecman.2015.05.004

Karvonen, M., & Kässi, T. (2013). Patent citations as a tool for analysing the early stages of convergence. *Technological Forecasting & Social Change, 80*(6), 1094-1107.

Kodama, F. (1992). Technology Fusion and the New R&D. *Harvard Business Review, 70*(4), 353–354.

Kranakis, E. (2013). Advances in Network Analysis and its Applications. *Mathematics in Industry, 18*.

Miller, R., Northup, T. E., Birnbaum, K. M., Boca, A., Boozer, A. D., & Kimble, H. J. (2005). Trapped atoms in cavity QED: coupling quantized light and matter. *Journal of Physics B Atomic Molecular & Optical Physics, 38*(9), S551-S565.

Nordmann, A., Bruland, K., Bibel, W., & Others, A. (2004). Converging Technologies – Shaping the Future of European Societies. *Interim Report of the Scenarios Group*.

Pavitt, K. (1984). Sectoral patterns of technical change: Towards a taxonomy and a theory. *Research Policy, 13*(6), 343-373.

Porter, A. L., Youtie, J., Shapira, P., & Schoeneck, D. J. (2008). Refining search terms

for nanotechnology. *Journal of Nanoparticle Research, 10*(5), 715-728.

Poudel, B., Hao, Q., Ma, Y., Minnich, A., Muto, A., Lan, Y. C., . . . Vashaee, D. (2008). High Thermoelectric Performance of Nanostructured Bismuth Antimony Telluride Bulk Alloys. *Science, 320*(5876), 634.

Roco, M. C., & Bainbridge, W. S. (2003). *Converging Technologies for Improving Human Performance*: Springer Netherlands.

Rotolo, D., Hicks, D., & Martin, B. R. (2015). What is an emerging technology? *Research Policy, 44*(10), 1827-1843.

Song, C. H., Elvers, D., & Leker, J. (2017). Anticipation of converging technology areas — A refined approach for the identification of attractive fields of innovation. *Technological Forecasting & Social Change, 116*, 98-115.

Tunzelmann, G. N. V. (1998). Localized Technological Search And Multi-Technology Companies. *Economics of Innovation & New Technology, 6*(2-3), 231-256.

Xu, S., Wang, X., Liu, Z., & Luan, C. (2013). Network structural analysis of technology: a study from patent perspective. *Journal of Science & Technology Policy in China, 4*(3), 214-235.

You, Y. B., Kim, B. K., & Jeong, E. S. (2014). An exploratory study on the development path of converging technologies using patent analysis: the case of nano biosensors. *Asian Journal of Technology Innovation, 22*(1), 100-113.

Scientific research collaboration in Artificial Intelligence: global trends and citations at the institution level

Lipeng Fan [1] Yuefen Wang [2] and Shengchun Ding [3]

[1] funnypower@126.com
Nanjing University of Science & Technology, School of Economics & Management, Nanjing(China)

[2] yuefen163@163.com
Nanjing University of Science & Technology, School of Economics & Management, Nanjing(China)
Jiangsu Collaborative Innovation Center of Social Safety Science and Technology, Nanjing(China)

[3] todingding@163.com
Nanjing University of Science & Technology, School of Economics & Management, Nanjing(China)

Abstract

In order to gain a deeper understanding of collaboration and the relationship between collaboration patterns and citations in global Artificial Intelligence (AI) research, this present paper defines institution types and collaboration patterns from a new perspective. According to a variation of H-index, it classifies institutions into two types: Main institutions and Normal institutions. Based on institution types of the first and remaining institutions in a paper, it divides collaboration publications into six parts: M, M&M, M&N, N, N&M and N&N. In this study, all publications were collected from papers listed in Web of science from 1997 to 2017, published in the field of AI. According the number of units in a paper, results show that five or more authors have a great chance to be the primary pattern in AI field in the future; single-institution papers are the primary pattern but decreasing sharply during a long time; single-country papers keep playing a dominant role in past almost 20 years. According to different collaboration types, results show that five or more author publications are the primary form in M&M, M&N and N&M types, while three-author papers in N&N; Domestic two-institution papers in M&N and N&N are obviously more than that in M&M and N&M types; Single-country papers account for a large share in M&N, N&N and N&M, while two-country papers are more than single-country papers and become the most important part since 2010 in M&M. According to the relationship between collaboration types and citations, results show that the number of Main institutions has a positive relationship with the citation values, while the number of Normal institutions has a little negative influence on N&N type.

Keywords

Collaboration pattern, Global trends, Citation, Institution level, Artificial Intelligence.

Introduction

Scientific research collaboration refers to the form between individuals and individuals, individuals and groups, groups and groups work together to accomplish the same scientific research task. Scientific research is a complex and arduous group work. The interaction between people in scientific research activities directly affects the completion of scientific research collaboration and programs. Through research collaboration, however, knowledge and culture can be promoted to exchange, new academic thoughts can be generated, the academic

communication network can be built, research productivity can be increased, and research costs can be decreased (Katz and Martin, 1997; Beaver, 2001; Lili Yuan, 2018). Some studies have revealed that collaboration papers are becoming more and more prevalent in scientific research (Abramo et al., 2004; Nguyen et al., 2017).

Many researchers have paid attention to the collaboration patterns from different dimensions. Sooryamoorthy (2009) classifies the collaboration patterns into domestic, international, intra-institutional and inter-institutional according to the number of countries or institutions in a publication. Liu et al. (2012) following the classification concept of patentees explore the institution collaboration papers among six levels: private enterprises, government, university, hospital, research institutes and not-for-profit organizations. Wang et al. (2017) investigate collaboration patterns from scholar's local perspectives based on their academic ages. Lee et al. (2012) propose four types of collaboration by categorizing network analyses into two dimensions: structural positions and the relational characteristics of individual nodes. Besides, there are numerous studies on the relationship between collaboration patterns and citations. Some of them have been performed to analyse the citations across the number of research units. The research of Gazni and Didegah (2011) has shown that there was a positive correlation between the number of citations and the number of authors and institutions. Ibáñez et al. (2013) have defined that international collaboration results on average in publications with higher citation rates than national and institutional collaborations. Also, some of them focus on the citation across the characters of research units. Based on the economic levels of countries, Ni and An (2018) analyse the relationship between the number of international collaboration papers and citations according to different international economic collaboration types. Zhang et al. (2018) examine the collaboration and citations based on the authors' attributes: productivity, impact, research interests, and gender. Although the existing studies about collaboration patterns and citations are abound, little attention has been paid to the interactions from the institution perspective. As the institution is the main driver research units in scientific research, it is important to explore the collaboration trends and the relationship between collaboration types and citations at the institution level.

In recent years, Artificial Intelligence (AI), as a branch of Computer Science, draws a great attention of scientists and becomes one of the most popular terms on internet. As an interdisciplinary, AI has attracted more and more researchers with respect to its theories and principles since the 1956 Dartmouth conference (Pham D.T, 1999; Mohd Ali J, 2015). Besides, some countries have spent a lot of money to fund AI-related researches, and encourage a plenty of firms and institutions to develop AI projects. Moreover, there are many institutions applying for constructing new department of AI School.

As some studies indicate that both collaboration trends and the relationship between cooperation and citation patterns differ from discipline to discipline (Franceschet, 2011; Gazni et al., 2012), we take the field of AI as an example to analyse the cooperation trends and citations among different collaboration types at the institution level in this paper. In particular, we will address the following research questions from the institution perspective: (1) what are the collaboration patterns in AI filed at the research units' levels? (2) what are the differences in the collaboration trends for different collaboration types? (3) what is the relationship between collaboration types and citations?

This paper is divided into two parts. The first part is the collaboration trends for different cooperation types in global AI research. The second part is the relationship between collaboration types and citations.

Data and Methodology

Data Sources and Pre-processing

Web of science is a well-established database and most frequently used for bibliometrics analysis of scientific research. Since there is a category named "Computer Science, Artificial Intelligence" in its core database, we used "WC = Computer Science, Artificial Intelligence" to search publications, and the retrieved data is limited to articles published from 1997 to 2017. The collection work came to an end on May 10, 2018. Eventually, the dataset used in this study covers 765491 academic articles and 72916 institutions. In order to analyze the collaboration patterns of institutions, the author's name (AU), authors' address information (C1), publish year (PY), cited frequency (Z9), WOS number (UT) were extracted from the dataset. After excluding the papers without C1 tags, finally, a total of 744108 records were obtained.

Using C1 tag, the country information and institution information were extracted to help cleaning the institution data. Due to some institutions with same names are in different countries, we clean the country data firstly. For example, UK has four or more regions, such as England, Scotland, Wales and Northern Ireland. In common sense, England represents UK, so we transform the other three regions into England. Although Hong Kong, Macau and Taiwan are regions of Peoples R China, most institutions belonging to Hong Kong and Macau have written the country name as Peoples R China. So, we take Taiwan as an independent region in this paper, and replace the other two regions with Peoples R China.

Compared with country information extraction, identifying the unique name of institutions is a complicated task (van Raan, 2005). After analysing the institution information, we clean the institutions with the following errors: (1) *Name has been changed*, such as "Beihang Univ" and "Beijing Univ Aeronaut & Astronaut". (2) *Different abbreviations*, such as "Swiss Fed Inst Technol" and "ETH". (3) *Different positions of "Univ"*, such as "Univ Washington" and "Washington Univ". (4) *Institutional departments*, such as both "INRIA Rhone Alpes" and "INRIA Sophia Antipolis" all belong to "INRIA". (5) *Different translation*, such as "Free Univ Brussels" and "Univ Libre Bruxelles" actually are the same institution.

We merge the institutions with the above problems, and set the right name with the higher number of publications.

Define institution types

H-index, proposed by Hirsch, is a method to assess the influence of scientists by publications and citations simultaneously, and it can also be extended to evaluate the performance of institutions, countries and journals. When calculating publications and citations of institutions, it is difficult to distribute the credits among ordered co-institutions. Since the order of institutions in a paper represents a certain order of importance, we choose Arithmetic counting, one of the most prominent counting methods, to calculate the h-index of institution, and name this method as AH-index. In this counting method, the credits are linearly distributed in decreasing order among several co-institutions according to the following formula (Trenchard, 1992; Van Hooydonk, 1997):

$$c_i^n = \frac{n-i+1}{\sum_{k=1}^{n} k} = \frac{2(\mathrm{n}+1-\mathrm{i})}{n(n+1)}, (\mathrm{i} \in \mathrm{n})$$

In this formula, i is the position of the institution, n is the number of institutions in a paper. We use the same formula to obtain the citations of each institutions. If there are some same institutions in a paper, the publications and citations of this institution will be calculated multiply times.

Although corresponding author and first author are equally important in many cases, the corresponding institutions are the same with first institutions in almost 96% papers of our dataset. So, we only take the order of institutions into consideration to reduce the workload in our study.

After sorting in descending order according to the AH-index of institutions, we define the institutions that sum publications of which are not less than half of our dataset as Main institutions, and the remain as Normal institutions. In this paper, the number of Main institutions is 562 with 50.64% of total papers. Table 1 lists the top 20 of Main institutions.

Table 1. Top 20 of Main institutions

Rank	*Institution*	*Country*	*Publications*	*Citations*	*AH-index*
1	Nanyang Technol Univ	Singapore	4112.32	56775.6	82.69
2	Carnegie Mellon Univ	USA	3451.47	58065.23	79.03
3	MIT	USA	2235.01	41417.67	73.66
4	Chinese Acad Sci	P R China	6880.77	47583.21	68.45
5	Stanford Univ	USA	1601.14	31656.08	63.41
6	Univ Calif Berkeley	USA	1466.07	64332.18	63.37
7	Hong Kong Polytech Univ	P R China	1814.32	29688.51	61.61
8	Univ So Calif	USA	2085.81	27150.27	61.55
9	Univ Illinois	USA	2347.08	42383.32	61.06
10	Swiss Fed Inst Technol	Switzerland	1541.89	31233.73	60.7
11	Natl Univ Singapore	Singapore	2571.57	33375.42	59.09
12	Univ Oxford	England	967.01	26182.3	55.1
13	Tsinghua Univ	P R China	5262.58	30471.49	54.43
14	Ecole Polytech Fed Lausanne	Switzerland	1191.3	21716.58	53.92
15	Chinese Univ Hong Kong	P R China	1967.11	23341.94	52.14
16	Univ Maryland	USA	1695.92	24391.16	51.8
17	Univ Granada	Spain	1524.34	21332.6	51.22
18	Southeast Univ	P R China	1826.21	18886.26	50.12
19	Washington Univ	USA	1427.22	17108.22	49.51
20	INRIA	France	1003.76	19706.27	48.07

Define collaboration types

According to the types of the first and remaining institutions in a paper, we divide the collaboration papers in our study into six types: M, M&M, M&N, N, N&M, N&N. Here, M

represents Main institutions, N represents Normal institutions. Specifically, M type represents collaboration among one Main institution; M&M type represents collaboration that the first institution is Main institution and the remaining institutions have at least one Main institution; M&N type represents collaboration with first Main institution and remaining all Normal institutions; N type represents collaboration among one Normal institution; N&M type represents collaboration that the first institution is Normal institution and the remaining institutions have at least one Main institution; N&N type represents collaboration with first Normal institution and remaining all Normal institutions.

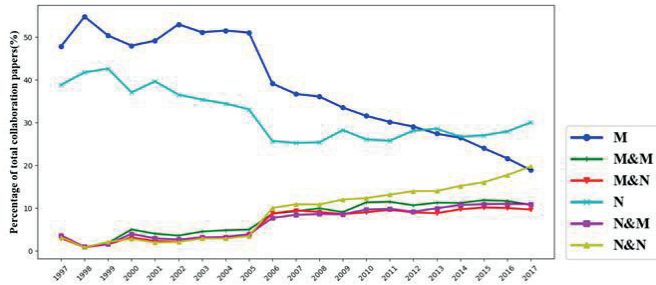


Fig 1. The trends of publications in six collaboration types

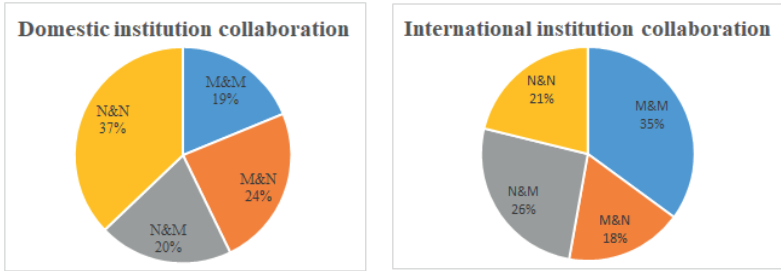


Fig 2. Proportion of domestic and international institution collaboration types

The total proportion of the six collaboration types is 36.03%, 7.87%, 6.74%, 32.72%, 7% and 9.64% respectively, and the trends of the proportion are shown in Fig 1. As illustrated in Fig 1, single-institution papers are the primary pattern in AI filed over the past almost 20 years, while the trend keeps descending. In single-institution papers, the proportion of M collaboration type is always higher than the other types before 2012, and after that, N collaboration type becomes the primary pattern until now. In multi-institutions papers, M&M collaboration type is the main pattern before 2005, and after that, N&N collaboration type rise quickly and surpass the other types. Meanwhile, M&N and N&M always keep rising shoulder and shoulder slowly. The results indicate that N and N&N collaboration types will become the mainstream in next few years, and the type of first institution has little effect on the number of publications in multi-institutions papers.

Fig 2 shows the differences in proportion of institution collaboration types between domestic and international publications. As we can see from Fig 2, N&N and M&N institution

collaboration types take a larger share in domestic institution collaboration publications, while that are M&M and N&M types in international institution collaboration papers.

Results and analysis

Collaboration trends of different collaboration types

In order to explore the collaboration trends in different collaboration types, the trends of publications at the author, institution, country level are investigated respectively.

Trends of collaboration patterns at the author level

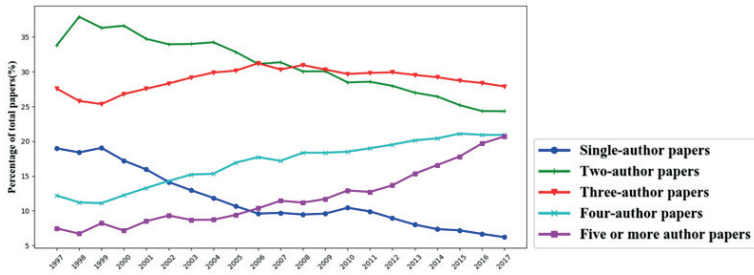


Fig 3. Trends of collaboration papers at the author level in AI research

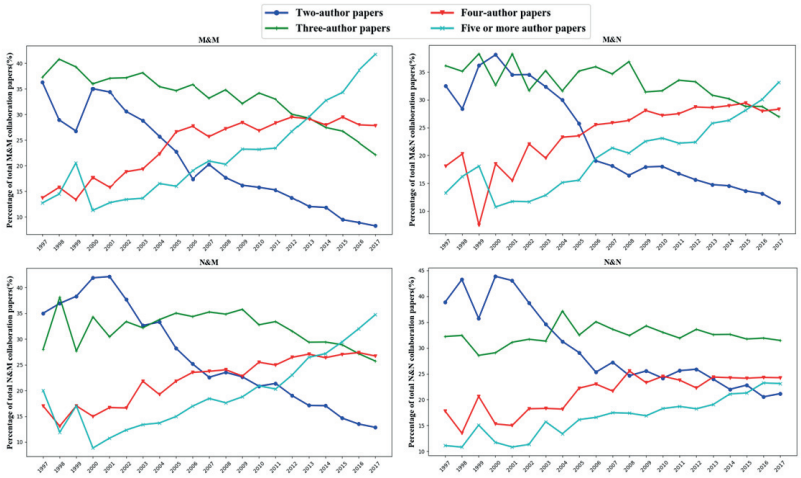


Fig 4. Trends of multi-institution collaboration types at the author level in AI research

Through quantitative analysis of AU tags, it is found that 89.9% of papers are collaborative publications in global AI field, which infers that collaboration papers are the mainstream in AI research.

In total dataset, trends of publications at the author level are shown in Fig 3. It displays that most of the lines all keep rising steadily except single-author and two-author papers. Although the trend of two-author papers inclines to downtrend dramatically with the single-author papers,

it is still the primary pattern at the author level in AI before 2008. The line of three-author papers has a noticeable rising during 1997-2017, and becomes the major part after 2008. Four-author and five or more author papers keep rising quickly, and the latter form seems to surpass the former one after 2017. This phenomenon indicates that the global AI research has a tendency toward three or more author publications, and five or more author papers have a great chance to become the new primary pattern in the future.

Besides, trends of multi-institution collaboration types are shown in Fig 4. In line with the previous studies, papers with at least two authors considered as collaboration works (Lariviere 2006). As we can see from Fig 4, in all four collaboration types, there is a dramatical downward trend in two-author papers and a slowly decline in the number of three-author papers, while five or more author papers keep rising and become more and more important. Fig 4 also reveals that the percentage of five or more author papers grows rapidly and already become the most important part in M&M pattern since 2013, and that happens in M&N and N&M patterns after 2015. For N&N type, three-author papers are always the main form since 2004, and the trends of four or more author papers keep increasing but slowly. All in all, the results show that five or more author publications are the primary form in M&M, M&N and N&M collaboration types, while that is three-author papers in N&N type.

Trends of collaboration patterns at the institution level

The percentage trends of different collaboration patterns at the institution level were investigated in global AI research during 1997-2017.

As illustrated in Fig 5, single-institution papers are the primary publishing pattern at the institution level, followed by two-institution papers. The number of three or more institution papers accounts for a small proportion of the total publications. However, the area of single-institution papers presents an obviously descending trend during 1997-2017, while multi-institution papers keep a steady increasing trend. Overall, single-institution papers will still have an absolute dominance compared with multi-institution papers in recent years. Besides, two-institution papers are the main form in both domestic and international institution collaboration papers, and four or more institution papers are more intend to international collaboration than domestic collaboration.

The trends of M&M, M&N, N&M and N&N types at the institutional level are shown in Fig 6. It illustrates that two-institution papers are always the primary pattern in all collaboration types, and only has a slowly down trend during the past almost 20 years. Domestic two-institution papers in M&N and N&N are obviously more than that in M&M and N&M types. The percentage of three or more institution papers are almost the same trend in M&M and N&M, and much more than that in M&N and N&N type. The results imply that there is probably a positive relationship between the number of Main institutions and three or more institution collaboration papers.

The trends of collaboration patterns at the country level

The percentage of domestic collaboration papers, international collaboration papers, single-country papers, two-country papers and three or more country papers were investigated in global AI during 1997-2017.

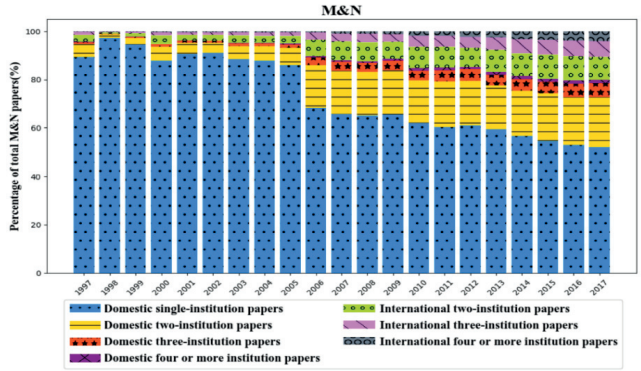


Fig 5. Trends of collaboration papers at the institution level in AI research

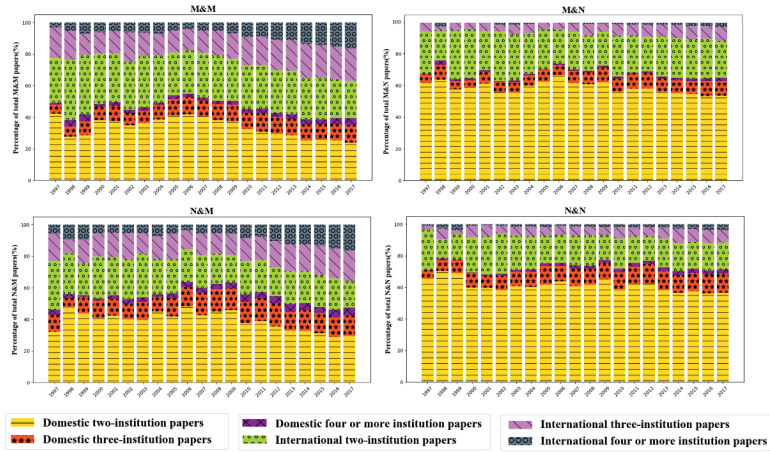


Fig 6. Trends of multi-institution collaboration types at the institution level in AI research

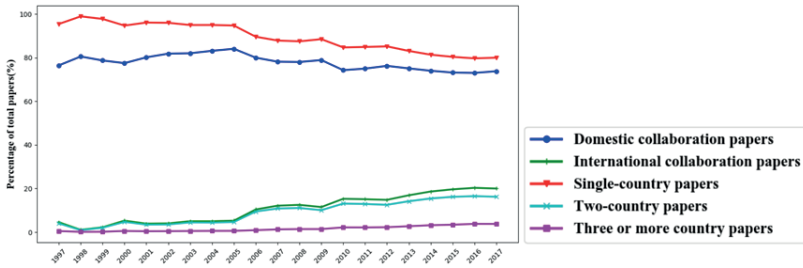


Fig 7. Trends of collaboration papers at the country level in AI research

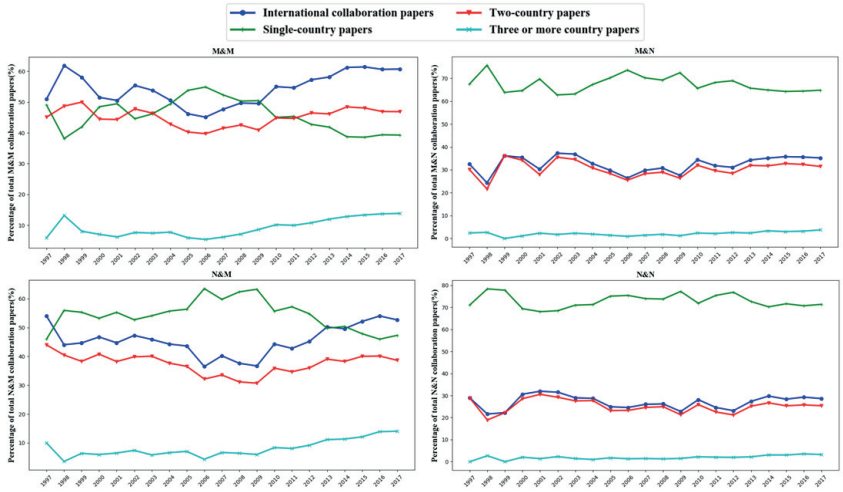


Fig 8. Trends of multi-institution collaboration types at the country level in AI research

As shown in Fig 7, at the country/region level, single-country papers play a dominant role in the past almost 20 years. Actually, the percentage of single-country papers keeps descending slowly from 100% to about 80%, while the multiple-country papers keep a noticeable upward trend to about 18%. Three or more country publications are also increasing, but the rate is much smaller. Fig 7 also shows that the trends of two-country papers and the international collaboration papers are shoulder by shoulder during 1997-2017. This indicates that two-country papers are the primary pattern of the international country collaboration in global AI research.

The trends of different collaboration types at the country level are shown in Fig 8, where single-country papers are the same with domestic papers. As we can see from Fig 8, single-country papers account for a large share in M&N, N&N and N&M, while in M&M, two-country papers are more than single-country papers and become the most important part since 2010. Besides, in the past almost 20 years, there are 16 years that international collaboration papers exceed single-country papers in M&M, while that only 4 years in N&M. Moreover, the average percentage of international collaboration papers in N&M and M&N types are higher than that in N&N. The results reveal that the type of first institution have a certain effect on the number of international collaboration papers.

Citations among different collaboration types and number of institutions

This section analyses the relationship between citations and collaboration papers at the institution level according to three aspects: collaboration types, number of institutions, and type of institutions.

Citations of different collaboration types

Table 2. Mean and standard deviation of citations for different collaboration types

Citation count	*Collaboration Types*					
	M	*M&M*	*M&N*	*N*	*N&M*	*N&N*

Mean	11.34	18.94	11.66	5.92	11.78	6.86
SD	24.82	36.1	21.71	10.23	20.99	11.45

Table 2 shows the average citations for different collaboration types in our study. From Table 2, we can see that the highest average number of citations (18.94 ± 36.1) corresponds to M&M papers, followed by N&M papers (11.78 ± 20.99), M&N papers (11.66 ± 21.71), M papers (11.34 ± 24.82), while the lowest average value corresponds to N papers (5.92 ± 10.23) and N&N papers (6.86 ± 11.45). The results indicate that the papers with Main institutions statistically have a significantly higher influence than that without Main institutions. Moreover, inter-institution types show a higher influence compared with intra-institution in collaboration papers with same institution type.

Citations of different number and type of institutions

Table 3 shows the average citations from the number and type of institutions perspective. As seen in Table 3, papers published with five or more Main institutions obtain the highest average value (26.31 ± 44.35). Besides, the average citations vary little with the increase of Normal institutions. The results indicate that the more the number of Main institutions in papers, the higher the average value of citations. Meanwhile, the number of Normal institutions has little to do with the average value of citations.

Table 3. Mean and standard deviation of citations according to the number and type of institutions

Institution Type	*Citation count*	*Number*					
		0	*1*	*2*	*3*	*4*	*>=5*
Main institution	Mean	5.64	10.89	17.33	20.41	23.43	26.31
	SD	8.53	21.09	31.90	34.50	40.28	44.35
Normal institution	Mean	12.2	7.66	7.67	8.05	7.87	10.9
	SD	24.41	12.91	11.71	12.61	12.3	22.28

Citations of collaboration types according to the number and type of institutions

Table 4. Mean ± standard deviation of citations for different collaboration types according to the number and type of institutions

Institution Type	*Number*	*Collaboration Types*			
		M&M	*M&N*	*N&M*	*N&N*
Main institution	0	-	-	-	6.34±9.5
	1	-	10.98±18.71	10.27±16.61	-
	2	17.72±32.95	-	14.08±22.63	-
	3	20.44±33.92	-	18.11±32.63	-
	4	23.69±40.37	-	18.31±28.78	-
	>=5	24.32±37.31	-	30.08±51.92	-
Normal institution	0	19.28±35.77	-	-	-
	1	16.22±27.69	11.07±19.06	11.53±19.28	-

	2	14.34±21.85	10.04±15.21	9.33±13.17	6.34±9.72
	3	14.08±19.87	11.34±19.9	9.99±15.97	6.05±8.16
	4	11.42±12.88	10.85±25.27	10.21±16.97	5.21±6.19
	>=5	15.09±20.69	5.25±7.3	13.51±24.94	5±5.62

We describe the evolution of different collaboration types in our study according to the number and type of institutions in Table 4. With the same number of institutions, most average values of M&M type are larger than that of N&M, and there is a noticeable downward of the average citations with the collaboration of M&M, M&N, N&M and N&N, indicating that it has a certain positive influence on citations that the first institution is a Main institution. According to the same collaboration types, we find that, the more the number of Main institutions is, the higher the average citations of M&M and N&M are. Moreover, there is no obvious relationship between the number of Normal institutions and the citation value in M&M, M&N and N&M types, whereas it has a little negative influence on N&N type.

Conclusion

Using the data of almost 20 years in AI field from Web of science, we take the institution influence into consideration and divide the dataset into six parts according to collaboration types at the institution level: M, M&M, M&N, N, N&M and N&N. From the number of authors, institutions and countries perspective, as well as citation, we explored the global trends and differences among the collaboration types.

On the whole, at the research units' levels, we find that five or more author papers keep rising quickly, and have a tend to become the main form in AI field. Moreover, single-institution papers are the primary pattern but decreasing sharply, and single-country papers keep playing a dominant role in past almost 20 years. According to different collaboration types, five or more author publications are the primary form in M&M, M&N and N&M types, while three-author papers in N&N. Domestic two-institution papers in M&N and N&N are obviously more than that in M&M and N&M types. Single-country papers account for a large share in M&N, N&N and N&M, while in M&M, two-country papers are more than single-country papers and become the most important part since 2010. From citation perspective, the number of Main institutions has a positive relationship with the citation values, while the number of Normal institutions has a little negative influence on N&N type, and has no obvious relationship with other collaboration types. Moreover, it has a certain positive influence on citations that the first institution is a Main institution.

Admittedly, our study has some limitations. First, we only choose AI data in our study, the results may not be generally applicable. Second, when analysing the relationship between citations and collaboration types, this paper does not consider the total number of institutions in each paper, which may have a certain impact on the results.

In the future work, we are interested in capturing the trends of institution collaboration types among dominant countries, and take more effective methods to explore the citations.

Acknowledgments

The authors are grateful to anonymous referees and editors for their invaluable and insightful comments, and thank for the support by the National Social Science of China (16ZDA224).

References

Abramo, G., D'Angelo, C. A., & Solazzi, M. (2004). The relationship between scientists' research performance and the degree of internationalization of their research. *Scientometrics*, 86(3), 629–643.

Beaver, D. D. (2001). Reflections on scientific collaboration (and its study): Past, present, and future. *Scientometrics*, 52(3), 365–377.

Franceschet, M. (2011). Collaboration in computer science: A network science approach. *Journal of the American Society for Information Science and Technology*, 62(10), 1992–2012.

Gazni, A., & Didegah, F. (2011). Investigating different types of research collaboration and citation impact: a case study of Harvard University's publications. *Scientometrics*, 87(2), 251-265.

Gazni, A., Sugimoto, C. R., & Didegah, F. (2012). Mapping world scientific collaboration: Authors, institutions, and countries. *Journal of the American Society for Information Science and Technology*, 63(2), 323–335

Ibáñez, A., Bielza, C., & Larrañaga, P. (2013). Relationship among research collaboration, number of documents and number of citations: a case study in Spanish computer science production in 2000–2009. *Scientometrics*, 95(2), 689-716.

Katz, J. S., & Martin, B. R. (1997). What is research collaboration? *Research Policy*, 26(1), 1–18.

Lariviere, V., Gingras, Y., & Archambault, E. (2006). Canadian collaboration networks: A comparative analysis of the natural sciences, social sciences and the humanities. *Scientometrics*, 68(3), 519–533.

Lee, D. H., Seo, I. W., Choe, H. C., & Kim, H. D. (2012). Collaboration network patterns and research performance: the case of Korean public research institutions. *Scientometrics*, 91(3), 925-942.

Liu, H. I., Chang, B. C., & Chen, K. C. (2012). Collaboration patterns of Taiwanese scientific publications in various research areas. *Scientometrics*, 92(1), 145-155.

Mohd Ali, J., Hussain, M.A., Tade, M.O., & Zhang, J (2015). Artificial intelligence techniques applied as estimator in chemical process systems—A literature survey. *Exp. Syst. Appl*, 42, 5915–5931.

Nguyen, T. V., Ho-Le, T. P., & Ut, V. Le. (2017). International collaboration in scientific research in Vietnam: An analysis of patterns and impact. *Scientometrics*, 110(2), 1035–1051.

Ni, P., & An, X. (2018). Relationship between international collaboration papers and their citations from an economic perspective. *Scientometrics*, 116(2), 863-877.

Pham, D.T., Pham, P.T.N (1999). Artificial intelligence in engineering. *Int. J. Mach. Tools Manuf*, 39, 937–949.

Sooryamoorthy, R. (2009). Do types of collaboration change citation? Collaboration and citation patterns of South African science publications. *Scientometrics*, 81(1), 177-193.

Van Raan, A. F. J. (2005). Fatal attraction: Conceptual and methodological problems in the ranking of universities by bibliometric methods. *Scientometrics*, 62(1), 133–143.

Wang, W., Yu, S., Bekele, T. M., Kong, X., & Xia, F. (2017). Scientific collaboration patterns vary with scholars' academic ages. *Scientometrics*, 112(1), 329-343.

Yuan, L., Hao, Y., Li, M., et al. Who are the international research collaboration partners for China? A novel data perspective based on NSFC grants. *Scientometrics*, 2018, 116(1):401-422.

Zhang, C., Bu, Y., Ding, Y., & Xu, J. (2018). Understanding scientific collaboration: Homophily, transitivity, and preferential attachment. *Journal of the Association for Information Science and Technology*, 69(1), 72-86.

Mobility of African doctoral graduates from South African universities

Michael Kahn[1], Thandi Gamedze and Joshua Oghenetega
School of Government
University of the Western Cape
Bellville 7535
South Africa

Introduction

This paper reports on the first stage of a project 'Mobility of the Highly Skilled' (MOTHS). The project has set out to understand the phenomena of brain drain, brain gain and brain circulation of students from sub-Saharan Africa that travel to South Africa to undertake doctoral study. The MOTHS Project is housed at the University of the Western Cape and employs a unique tracer method to obtain data on the career paths of these students. The project is relevant to labour policy as high-level skills are deemed inadequate for the country to engage with existing knowledge economy imperatives and the emerging challenges of the fourth industrial revolution (DST, 2019).

Twenty-five years ago, as enshrined in its Bill of Rights and an inclusive constitutional democracy, South Africa transitioned from apartheid repression to equal opportunity. By design, apartheid had restricted quality schooling and post-school education opportunities of the African majority, thereby reducing the talent pool available for personal and national development. In addition, the country was isolated from worldwide cultural, political and technological exchanges. The inequities of that period manifest into the present, being evident in general high-skills shortages and demographic skewing in management, academia, and all levels of skill. While university student and staff demographic profiles have shifted considerably, with African students now in the majority, this tails off at postgraduate level, especially for doctoral study.

Opening up of the country has seen large numbers of foreign African students enrolling in its universities, making up some 7% of all full-time enrolments. This opening up is vested in policy through conformance with the Southern African Development Community (SADC) Protocol on Education and Training that requires member states to set aside 5% of university places for students from other SADC countries, and for these persons to pay domestic fees for their study. The proportion of foreign students rises with postgraduate levels, so much so that foreign Africans comprise a third of South Africa's total doctoral graduates, outnumbering citizen African doctoral graduates by a factor of four (Cloete, 2015). At post-doctoral level the proportion of foreign African students is now 60% (DST, 2018).

Faced with ongoing skills shortage in the country, it might be anticipated that foreign graduates be induced to remain after graduation, the more so as the costs for all university students, at all levels of study, are subsidised by the SA government. However, it remains policy that upon graduation, foreign students should return to their country of origin. This provision has been justified as part of South Africa's commitment to the development of Africa north of the Limpopo River. Recommendations to introduce a quota for postgraduate work permits (DST, 2012; Presidency, 2012) have received no political support, and immigration of the highly skilled remains subject to burdensome compliance with the prescripts of a scarce skills scoreboard (RSA, 2014).

[1] Corresponding author: mjkahn@uwc.ac.za

Currently there is no information regarding the mobility of such foreign doctoral graduates. The main objective of the study is thus to fill the information gap: on graduation, where do these highly-skilled persons go? Do they return home and settle down? Do they join the African brain drain and migrate to North America or to Europe (OECD, 2016)? What proportion stay on in South Africa, and why? No official information is available to illuminate these questions, nor do the South African host universities possess such information. A tracer survey of the doctoral graduates for the five years 2012-2016 was thus designed, implemented and analysed to begin to answer these questions.

The paper commences with a review of the literature on the methods used to quantify mobility as expressed in brain drain, brain circulation and brain gain. The second section describes South Africa higher education enrolment and graduation patterns, and her role as a regional hub for postgraduate higher education. The third section presents the methodology of the MOTHS research project. The fourth section provides the results, followed by a concluding a discussion, with suggestions regarding the next steps.

Mobility of the Highly Skilled

The broad topic of migration is served by a substantial literature dealing with issues such as forced migration, state failure, violence, and remittance cultures. The literature covering the mobility of the highly skilled is less developed.

Internationally, highly skilled individuals, generally those with tertiary level qualifications, are seen as essential for the processes of innovation, scientific discovery, and knowledge generation. Host countries benefit significantly from the skill, knowledge and expertise that these individuals bring (Solimano, 2004; Moed and Halevi, 2014). As a result countries compete for the highly skilled through selective immigration policies to make themselves attractive destinations (Pekkala et al, 2016; Docquier and Machado, 2016). Strategies for attraction include elimination of tuition fees (Lalic et al, 2012), offers of scholarships or financial aid (UNESCO, 2013), the awarding of permanent residency to international students upon graduation (Group of Eight, 2014), points-based immigration policies (Mavrodi and Warren, 2013), tax benefits (Hercog, 2008), and easy access to work permits for foreign students in science, technology, engineering and mathematics (STEM) fields (Mugimu, 2010).

The positive impacts of the highly skilled often continue after they leave the host country, through their new relationships translating into knowledge flows (Agrawal, 2014). These individuals act as a connection between countries, contributing positively to their social, cultural, political and commercial relations (Levent, 2016; Go8, 2014).

Unfortunately, detailed assessment of the mobility of the highly skilled has proven to be difficult, with research limited by the coarse granularity of administrative data acquired at ports of entry, national census databases, and household surveys. A notable effort to enhance the quality of information on mobility was the OECD/UNESCO/Eurostat project 'Careers of Doctorate Holders' (CDH) that set out to understand graduate labour market outcomes, their career paths and mobility (Auriol, Misu and Freeman, 2013). CDH recognized the difficulties that statistical agencies would face in compiling a register of doctoral holders, and thus accepted that there would be diverse approaches in developing a sample frame. A sample fraction of 20% was set as the desirable minimum, with responses to be obtained by post, online, or through telephonic interview. Twenty-five OECD member states participated in the first round of the CDH survey. The main objective of CDH was to shed light on the value-add of doctoral studies, namely economic mobility, and not specifically migration characteristics

per se. Descriptive statistical methods were applied to the survey data in the respective country reports.

Other scholars of mobility have made use of interviews and surveys (Conchi and Michels, 2014), while Anas and Wickremasinghe (2010) tracked down a subsection of the highly skilled Sri Lankan diaspora through their National Science Foundation's expatriate scientist database. The German National Library hosts a structured dissertation database linked to detailed metadata at each university, allowing for a detailed search for graduate information, with the prospect of conducting direct surveys.

Over the last decade, new approaches to track mobility have been developed using bibliographic data on researchers. Such studies are only applicable to track the scientifically-active, whose publication outputs may be mined to track their movement from declared institution to institution (Moed and Halevi, 2014; Moed and Plume, 2013, Conchi and Michels, 2014). This method has produced useful information on the relationship between mobility and publication outputs and quality(see e.g. World Bank/Elsevier, 2014). Conchi and Michels (2014) discuss the limitations of bibliometric approaches compared with in-depth surveys. It is obvious that such surveys, whilst powerful in and of themselves, are limited to persons who are publishing in journals that are indexed to the major bibliographic databases such as Web of Science, Scopus, the International Bibliography of the Social Sciences and so on.

The South African Context

South Africa has for long been a hub for immigrants from Africa, and beyond. The late 19th century onset of industrial scale mining saw influxes of skilled and unskilled workers, the former from Cornwall and California; the latter from as far afield as the Belgian Congo. In the first half of the 20th century this influx of manual workers was supplemented by a small number of Africans seeking quality schooling and higher education. These hopefuls included future African presidents such as Seretse Khama, Robert Mugabe and Julius Nyerere. However, with the implementation of apartheid, such opportunities for both local and foreign Africans were terminated (Hartshorne, 1990). This loss of opportunity is evidenced in the fact that in 1994, when 'persons of colour' comprised some 80% of the population, less than 5% of graduate staff of the public research organisations were 'black' (LHA consultants, 1994). Post-apartheid redress was therefore a moral, political and economic necessity, with skills mobility into and out of the country a matter of policy concern.

Prior studies on the mobility of the highly skilled have received attention from local scholars, notably Meyer and Brown (1999), Kahn et al (2004), Kaplan (2008), Erasmus and Breier (2009), and Höppli (2014). All of these studies consider the macro level, being largely restricted to the use of foreign recipient country immigration databases and other secondary data. The studies serve to quantify the scale of emigration, be this short-term or permanent. In 2013 some 750 000 South African born individuals (the majority of whom are highly-skilled) resided in 23 major destination countries (Hoppli, 2014; Kaplan and Hoppli, 2017). By design, such studies are unable to shed any light on the demography of the migrants, save for the fact that they originated from South Africa.

This loss of skill points to perceived opportunities abroad. Despite exhortation to grow the innovation system (DST, 2002) the research staff headcount has remained static, increasing from 18 879 in 2005 to 19 217 in 2015 (DST, 2007; 2017). In contrast the number of doctoral and post-doctoral students increased from 10 002 to 22 422, implying a massive increase in supervisor workload (DST, 2007; 2017. The Frascati Manual guidelines (OECD, 2015) stipulate that doctoral and post-doctoral students are a component of the total stock of researchers. The total stock of researchers has therefore increased. It is moot whether this

increase, with its natural churn rate, is sustainable or effective in building deep research capacity.

Post-graduate enrolments demographics are becoming more representative of the national demography, with 2012 being the first year in which the total cohort of doctoral graduates from SA universities included more 'black' students than white students. However, closer inspection shows that a large percentage of the 'black' cohort are not South African nationals but foreign students (Cloete et al, 2015). This attests to the fact that South Africa has once again become a destination for migrant students from Africa, on a far larger scale than before apartheid.

Of the national total of 9000 doctoral graduates for 2012-2016, some 3300 foreign graduates hailed from thirty-four sub-Saharan countries (DHET, 2018). South African universities thus function as a postgraduate education hub for Sub-Saharan Africa. The post-graduation destination of these doctorate holders is unknown.

Methodology

The dearth of fine-grained data on the movement of the highly skilled into and out of South Africa has been noted above. Moreover, since the South African Ministry of Home Affairs does not capture emigration data, such administrative data as has been interrogated draws on foreign sources. Even if these sources are complete, the underlying databases are constructed to serve destination country information requirements, and can provide little information to inform South African policy.

It is of course possible to apply bibliometric methods to track the movement of academics over time, by capturing their changing institutional attribution in their publications, provided these are indexed on a major bibliographic database. This method only applies to academics who are publishing in the indexed literature, and is unlikely to shed much light on recent doctoral graduates who may not yet have developed a publication record. Accordingly, a different approach is required to quantify the geographic mobility of doctoral graduates.

Following the approach of the OECD/Eurostat Careers of Doctoral Holders project, the MOTHS Project set out to conduct a direct survey of graduates to determine their individual career paths. The survey instrument seeks detailed information from each graduate, including name, nationality, prior education and employment, host university, year of enrolment and graduation, dissertation title, field of science, information on funding, subsequent employment, and their experiences in the host university.

The success of the survey pivots on the availability of a complete, accurate and accessible national register of doctoral graduates. In the case of South Africa no such national register exists.

Furthermore, the Department of Home Affairs, that is responsible for immigration and emigration procedures, restricts access to visa, study and work permit information.

A possible source of information lies within the main granting agency for postgraduate studentships, the National Research Foundation (NRF). However the NRF extends grants to but 30% of all postgraduate students, so that for the purposes of the survey its studentship database is effectively incomplete (NRF, 2017).

As regards the universities, it is to be expected that their administration offices would maintain student registers that could be accessed for legitimate research purposes. Such registers are maintained behind a number of confidentiality firewalls, the most recent such being compliance with the prescripts of the Protection of Personal Information Act No.4 of 2013. University administrators exercise considerable reserve in allowing researcher access to student

information. Attempts to obtain direct access to student alumni databases were essentially fruitless.

Another possible source of information is that of graduation ceremony programmes that in some cases include biographic information of the doctoral awardees. Unfortunately, publication of such biographies differs among the universities – some disseminate biographies, whilst others do not. And even where the biographies are available, there is no uniformity of style. Worse still, graduation programmes are not available online, but are generally only obtained as a hard copy at a graduation ceremony. Building a register based on graduation lists would literally require attendance at each graduation ceremony, or personal study of the archives at each university of interest. Obviously, the absence of archival sources would imply that only current graduate lists could be compiled. Even the recent past would be unquantifiable.

The most promising source for constructing a register would appear to be that of the Department of Higher Education and Training, whose Higher Education Management Information System (HEMIS) database receives detailed data from the twenty-six public universities. Each public university is required to submit a structured, detailed and auditable annual return to the Department, that then uses this data to populate the input and output variables of the statutory budget allocation formula. Individual fields capture data on staff, students, subjects of study, type and level of study, nationality, group, gender, year of enrolment and completion, and publication outputs, including journal articles and books. Unfortunately, even anonymized HEMIS microdata was not available to this study.

This left but one untapped source that might serve to provide biographic information, namely official dissertation repositories. As in many other countries, it is a legal requirement that a pdf copy of a postgraduate dissertation is deposited with the library of the awarding institution. Each university library homepage has a link to their dissertation repository, and the National Research Foundation assembles all such university repositories into its Electronic Theses and Dissertations Portal.

At face value therefore, repository databases could assist the researcher interested in compiling graduate information. Unfortunately, the repository databases are unstructured, and host all dissertations and research reports for a given year, including all masters and doctoral degrees. The repositories are not searchable by category as they are not organized according to a standard specification by field type. In particular, nationality is not recorded in the pdf submissions.

Accordingly, the dissertation records for a given university, and given year, must be inspected, record by record, to identify doctoral dissertations. Thereafter some means of determining the nationality of a doctoral graduate must be found.

In a few cases the dissertation author declares 'acknowledgements' that are detailed enough to reveal nationality, and sometimes their undergraduate university and even a current employer. But in most cases, the nationality of a graduate can only be surmised at by an investigator who is familiar with the country of origin of African surnames.

Algorithms linking surname and country of origin are under development (see e.g. Rachevsky and Pu, 2011) but were not readily available for the type of research here envisioned.

It was therefore necessary to develop an accurate, valid, verifiable and repeatable method to extract nationality information from the dissertation repositories. The research team performed this extraction by direct inspection and triangulation.

The first author, a South African, with many years of work experience elsewhere in Africa, therefore conducted an experiment to identify foreign African doctoral graduates by direct inspection of each repository record. This was effected for the year 2012 for two leading research universities. A list of likely 'foreign African' doctoral graduates was compiled.

This immediately raised the question of the validity of the inspection exercise, namely what proportion of 'foreign Africans' were correctly identified. A simple triangulation check of the method was undertaken – the first author and research assistant (a sub-Saharan African masters student) separately examined two thousand dissertation records to extract 'foreign African' doctoral graduates.

Blind comparison of the respective lists, showed 90% coincidence in their identification of the doctoral graduates.

An additional check was provided by comparing the university totals for the year 2012 with the number of foreign African graduates recorded to the Department of Higher Education and Training database. The list totals showed 90% coverage, suggesting that the method would be robust enough to proceed further.

Results

The sampling strategy, the 'Repository Approach' was then applied to the dissertation repositories of the five leading research universities: Cape Town, Stellenbosch, Witwatersrand, Pretoria, and Kwazulu-Natal for the years 2012 and 2013.

The Department of Higher Education and Training database count of sub-Saharan African doctoral graduates without permanent resident status for the five universities over 2012 to 2013 totalled 494. With this target in mind, the Repository Approach was applied to the visual inspection of 10,000 individual records. A total of 420 potential respondents, or 85% of the total, were so identified.

Social media search tools including Google™, Google Scholar™, LinkedIn™ and Facebook™ were then used to match potential respondents with their email address. In this way a respondent database was populated with contact and other information, with email addresses identified for 360 doctoral holders.

The sample of 360 potential respondents were then approached by means of a personalized email to participate in the survey. A total of 336 emails were successfully despatched. A follow-up reminder email was sent after two weeks. The eventual realized sample amounted to 104 responses.

The realized sample was then tested for representivity against the frame of 494 graduates. Given the focus on mobility, representivity by country was the most important criterion. The Pearson chi test was applied to the frame and sample, and was found to be significant with $p < 0,05$. Sample representivity by country showed no significant difference.

As expected, the largest group of graduates were from Zimbabwe (23%), followed by Nigeria (21%). Tests of representivity by university of graduation, or field of study must be deferred to future analysis of a larger realized sample.

The sample characteristics are summarized as Table 1.

Table 1: Sample characteristics

Attribute	Value
Frame	494
Repository Approach sub-Frame	420
Sample	336
Respondents	104
Return rate (%)	31
Nationality: chi-squared	$p < 0,05$

Table 2: Survey Results

Attribute	Value
Median age (years)	40
Returnees (brain circulation) (%)	69
Out of Africa (brain drain) (%)	8
Employed in SA (brain gain) (%)	23
Employed at awarding university (%)	6
Funded by SA (%)	44
Self-funded (%)	19
External financial support (%)	28
Home university employment (%)	52

The realized sample has median age of 40 years at graduation (Table 2). This is in accord with that for all doctoral candidates (Mouton, 2011). As with their local peers, the foreign students are essentially middle-aged.

Unlike many of their local peers, especially those in the social sciences and humanities, they are in full time study, and mostly complete their degrees in the minimum statutory period of three years. Local graduates, a mix of full-time and part-time students, require an average of 4,5 years to complete their studies.

A second feature of the realized sample is the 24% representation of women. This may be compared with the average level of 40% females among all doctoral graduates (Mouton, 2011).

Regarding mobility, the data shows that 80 graduates (77%) had exited South Africa; while 24 (23%) remained. The group who exited comprised of two sub-groups: those who returned home, as a product of brain circulation, and those who exited Africa entirely, as part of continental brain drain. The brain circulation group made up 69% of the realized sample. The brain drain out of Africa amounted to 8% Of the brain gain group who remained in South Africa, three graduates were employed in a public research organization; the remainder were in the employ of a university. One third of the brain gain group who were in academia were employed at the university that had awarded their doctorate.

The doctoral graduates, irrespective of their being part of brain gain, circulation, or drain, are predominantly employed in higher education. This is consistent with the findings of Auriol et al (2013) for the OECD Careers of Doctorate Holders study, which showed that most PhD graduates end up working within higher education.

The largest sub-group of graduates (44%) enjoyed South African financial support. The next largest sub-group (28%) received foreign financial support from employers, their government, or donors. Finally, 19% of foreign African graduates made use of loans or their own funds.

Discussion

It has been shown that mining dissertation repositories coupled with social media is an efficient means of generating a database of potential respondents to an online survey. The first phase of the research suggests that South Africa has succeeded in growing high skilled capacity in sub-Saharan Africa.

As evidenced by the various incentives that countries employ, global competition to attract doctorate holders is high, yet for the group of sub-Saharan African PhD graduates, the home pull factor appears to be more compelling.

The low proportion of women graduates found in this study differs from the global pattern reported in Pekkala et al (2016) who show that for close to a decade the numbers of highly skilled female migrants have been greater than those of their male counterparts.

On the other hand, Habti (2012) finds that the careers of the female highly skilled are more likely to be affected by their roles within their families. This may well be the case for the MOTHS project set. Even the limited information gleaned from inspection of the dissertation acknowledgements shows mention of a family at home, with thanks for support expressed toward their spouses. This suggests an explanation why the majority of doctoral holders returned home: they re-join their family and re-assume employment. Further inquiry will be needed to understand the gender dimension in mobility.

As to the completion period for the doctoral degree, it is in the interests of the migrant students, their employer at home, and the host university, that they complete their degree in the minimum prescribed three-year period. Efficient completion reduces personal costs, and is also of benefit to the host university which receives the state student subsidy grant with minimum delay.

South Africa's role as a sub-Saharan PhD hub is not without cost to the country. Financial support for doctoral graduates flows through National Research Foundation funding of the South African Research Chairs Initiative and the Centres of Excellence programme, both of which enable the Research Chair or Centre of Excellence head to issue PhD studentships. Other financial support takes the form of the state student fee subsidy and university research grants to the student.

With just 23% remaining in the country after graduation, and making a contribution to South Africa's R&D activities, the direct, long-term returns to South Africa are low. This is because the data suggest that even for this 'brain gain' group, a number are in short-term post-doctoral positions, so that their medium term residency future is uncertain.

Of course the foreign doctoral graduates benefit the host university in other ways. Firstly, during their studies some contribute to university publication outputs as co-authors with their mentors. This is of academic value to themselves and their prior and emerging knowledge networks, and a financial benefit to the host South African university that is entitled to claim the normal publication grant from the Ministry of Higher Education and Training. Not only do the students bring a wealth of experience to the host university, they also provide connection to the foreign research networks of which they are part. In some cases these individuals maintain academic contact with their South African host universities post-graduation. Bibliometric studies are required to ascertain the strength and duration of such networking.

The South African National Planning Commission has recommended that by 2030, seventy percent of academic research and instruction staff should hold a doctoral degree. In 2012, 34%

of academic staff were so qualified (CHET, 2012). Accordingly, active recruitment from the stock of foreign, and experienced African doctoral graduates could raise the profile.

Lastly, if the goal underlying the investment in foreign African students is to develop regional R&D capacity, and South Africa sees this process as an investment into the graduates' countries of origin, then the evidence of the high rate of brain circulation make this a worthwhile investment. The investment constitutes a form of quiet diplomacy for building advanced academic capacity across Africa. Building research capacity among the African partners of the to-be Square Kilometre Array telescope with its cross-border outposts is a deliberate goal of the Department of Science and Technology. This effort also requires further study.

The MOTHS project set out to understand the characteristics of foreign African students that join South African universities to study for their doctorates. It has broken new ground in showing how even unstructured repositories may be used to populate a graduate register that may be used to conduct a reliable tracer study. Recognizing the limitations of the realized sample, the results of this first phase serve to demonstrate the effectiveness of the tracer method, and provide insight into the educational experiences of doctoral graduates.

It is clear that the Repository Method may be extended to examine the post-PhD career trajectories of other groups, such as local African students, or students of other designated groups. And there is also the possibility of studying the masters cohort. Tenacity plus the power of social media has been demonstrated to achieve the desired outcome

The next phase of the project will be to cover the years 2014-2016 for the same five research universities. It is expected that the realized sample will then be large enough for a more detailed analysis to be carried out. In particular, the survey instrument includes two items requesting open-ended responses, namely 'What was the most positive aspect of your educational experience?' and 'What was the most negative aspect of your educational experience?' A further sixteen items are populated with predetermined responses from drop-down menus. A database of categorical and qualitative information will thus be available for regression analysis.

In order to allow evidence-based policy making fully to inform immigration and higher education policy, it is essential to construct an accurate, complete and maintainable national graduate register that may be interrogated under appropriate rules of access. This would entail a constructive dialogue among the Departments of Science and Technology, Home Affairs, and Higher Education and Training.

Appreciation is due to the graduates who participated in the MOTHS Study, to university administration staff, to the Department of Higher Education and Training, and staff of the National Research Foundation. The suggestions and insights of the independent reviewers of the paper are valued. Thanks are also due to Professor Jonny Myers for his careful critique.

The provision of grant 98773 from the National Research Foundation is duly acknowledged.

References

Agrawal, A. (2014). *Diaspora networks, knowledge flows and brain drain* (No. 15). Geneva: World Intellectual Property Organization.

Anas, M. U. M., & Wickremasinghe, S. I. (2010). Brain drain of the scientific community of developing countries: the case of Sri Lanka. *Science and Public Policy*, *37*(5), 381-388.

ASSAf (2010). Academy of Science of South Africa. 2010. The PhD study: An evidence-based study on how to meet the demands for high-level skills in an emerging economy. Pretoria, ASSAf.

Auriol, L., M. Misu and R. A. Freeman (2013), 'Careers of Doctorate Holders: Analysis of Labour Market and Mobility Indicators', OECD Science, Technology and Industry Working Papers, 2013/04, Paris: OECD Publishing. http://dx.doi.org/10.1787/5k43nxgs289w-en

Breetzke G D and D W Hedding (2017). The changing racial profile of academic staff at South African higher education institutions (HEIs), 2005–2013. Pretoria: University of South Africa.

Centre for Higher Education Transformation (CHET) 2012. 2013 South African public higher education key statistics. Cape Town: CHET.

Cloete, N., Sheppard, C., & Bailey, T. (2015). South Africa as a Phd Hub in Africa? *Knowledge Production and Contradictory Functions in African Higher Education*, *1*, 75.

Conchi, S., & Michels, C. (2014). *Scientific mobility: An analysis of Germany, Austria, France and Great Britain* (No. 41). Fraunhofer ISI Discussion Papers Innovation Systems and Policy Analysis. Karlsruhe: Fraunhofer ISI.

Department of Home Affairs (2017). The White Paper on International Migration for South Africa. Pretoria: Department of Home Affairs.

Department of Science and Technology (2012). Final Report of the Ministerial Review Committee on the Science, Technology and Innovation Landscape of South Africa. Pretoria: Department of Science and Technology.

Docquier, F., & Machado, J. (2016). Global competition for attracting talents and the world economy. *The World Economy*, *39*(4), 530-542.

DST (2007) National Survey of the Inputs to Research and Development. Pretoria: Department of Science and Technology.

DST (2012) Report of the Ministerial Review of the STI Landscape. Pretoria: Department of Science and Technology.

DST (2017) National survey of the inputs to Research and Development. Pretoria: Department of Science and Technology.

DST (2019). *White Paper on Science, Technology and Innovation.* Pretoria: Department of Science and Technology.

Erasmus, J., & Breier, M. (2009). Skills shortage in South Africa: case studies of key professions.

Group of Eight Australia (Firm)(Go8). (2014). International students in higher education and their role in the Australian economy.

Hartshorne, K. (1990). Post-apartheid education: a concept in process. *unpublished paper, June*.

Hercog, M. (2008). The role of the state in attracting highly-skilled migrants: The case of the Netherlands. *EIPAScope*, *2008*(3), 1-6.

Höppli, T. (2014). Is the brain drain really reversing? New evidence. *Policy Research on International Services and Manufacturing Working Paper*, *1*.

Kahn, M. (2015). Mobility of the highly skilled—Towards a non-racial South Africa. *Science, Technology and Society*, *20*(3), 369-388.

Kahn, M., Blankley, W., Maharajh, R., Pogue, T. E., Reddy, V., Cele, G., & Du Toit, M. (2004). Flight of the Flamingos: A Study on the Mobility of R&D Workers.

Kaplan, D. (2008). Science and technology policy in South Africa: Past performance and proposals for the future. *Science, Technology and Society*, *13*(1), 95-122.

Kaplan, D., & Höppli, T. (2017). The South African brain drain: An empirical assessment. *Development Southern Africa*, 1-18.

Lalic, A., Simonovska, V. & Von Estorff, U. (2012). The concept of internationalisation and the inevitability of mobility of highly skilled employees: What can the nuclear energy

sector in Europe learn from it? Joint Research Centre Science and Policy Reports. Brussels: European Commission.
Levent, F. (2016). The economic impacts of international student mobility in the globalization process. *Journal of Human Sciences*, *13*(3), 3853-3870.
LHA Consultants (1994). The employment composition and affirmative action programmes of the science councils in South Africa, Johannesburg: LHA Consultants (Mimeo).
Mavroudi, E., & Warren, A. (2013). Highly skilled migration and the negotiation of immigration policy: Non-EEA postgraduate students and academic staff at English universities. *Geoforum*, *44*, 261-270.
Meyer, J. B., & Brown, M. (1999). *Scientific diasporas: A new approach to the brain drain*. Management of Social Transformations (MOST) Programme.
Moed, H. F., & Halevi, G. (2014). A bibliometric approach to tracking international scientific migration. *Scientometrics*, *101*(3), 1987-2001.
Moed, H. F., & Plume, A. (2013). Studying scientific migration in Scopus. *Scientometrics*, *94*(3), 929-942.
National Planning Commission (2012). *National Development Plan: Vision 2030*. Pretoria: The Presidency.
NRF (2017). *National Research Foundation Annual Report 2015/16*. Pretoria: National Research Foundatiopn.
OECD (2015). OECD 2015. The Measurement of Scientific, Technological and Innovation Activities. Frascati Manual 2015. Guidelines for Collecting and Reporting Data on Research and Experimental Development. Paris, OECD.
OECD (2016). *OECD Economic Outlook 2016.* Paris, OECD.
Pekkala, S., Kerr, W. R., Özden, C., & Parsons, C. (2016). *Global Talent Flows* (No. 6203). CESifo Working Paper.
Rachevsky L and Pu, K Q (2011). Selection of features for surname classification. 2011 IEEE Conference on Information Reuse and Integration. DOI: 10.1109/IRI.2011.6009513
Republic of South Africa (2013). Protection of Personal Information Act. Act 4 of 2013. Pretoria: Government Printer.
Republic of South Africa (2014). Call for comments on the National Scarce Skills List: top 100 occupations in demand, Government Gazette 37678. Pretoria: Government Printer.
Solimano, A. and Pollack M. (2004). *International mobility of the highly skilled: The case between Europe and Latin America* (No. 1). Paris: OECD Publishing.
World Bank/Elsevier (2014). A Decade of Development in Sub-Saharan African Science, Technology, Engineering & Mathematics Research. Washington: World Bank

Can Bradford's law be applied to determine core subject terms in a subject domain?

Omwoyo Bosire Onyancha[1] and Dennis N Ocholla[2]

[1] *onyanob@unisa.ac.za*
University of South Africa, Department of Information Science, PO Box 392 Unisa 0003

[2] *ocholladn@unizulu.ac.za*
University of Zululand, Faculty of Arts, Private Bag x3886 Kwa-Dlangezwa

Abstract

This study attempts to use the principles of Bradford's law to determine the core concepts of information and communication technologies (ICTs) research within the information and knowledge management (IKM) research. Data were obtained from EBSCO Discovery Service's, Library Information Science and Technology Abstracts (LISTA) and the Library and Information Science Source (LISS), using a variety of keywords as search terms. The procedures for conducting a Bradford analysis were followed to determine the core ICTS subject terms within IKM research published between 1998 and 2017. The results indicate that the core subjects varied from one study period to another; multi-disciplinarity of subject terms was highly visible; and the dispersion of subject terms fits Bradford's law of dispersion. However, the dispersion of articles according to Bradford's zones is not accurately representative of the principles of this law. The study has implications for curriculum development, thesaurus construction, literature research, collection development, development of an information system and subject organisation and description. The findings have implications on the application of Bradford's law beyond the analysis of core journals or publications in a field.

Introduction

One of the areas that have received considerable attention from bibliometricians are studies seeking to determine core areas of study, research, competencies, subject terms, journals, and words in a text, among others, in a subject domain/field. Diverse methods and techniques have been employed to make such determinations. Among the most common are

- the core/periphery model, mostly used to identify core terms in a text, journal or subject area (e.g. Borgatti & Everett 1999; Onyancha & Ocholla 2009; Ocholla, Britz & Onyancha 2011; Onyancha & Mokwatlo 2012)
- the subject specialisation index, mostly used to assess the concentration of research in given topics, fields or disciplines (e.g. UNESCO 2005; Davis 1983; Pouris & Pouris 2009; Onyancha 2018)
- Bradford's law of scattering, commonly used to identify core journals in a subject area or field (e.g. Petersohn 2014; Singh & Bebi 2014)

This study focuses on the possible use of Bradford's law to determine the core subject terms in the subject domain of information and communication technologies (ICTs) within the context of information and knowledge management (IKM). Bradford's law, formulated in 1934 by Samuel C Bradford, states that "if scientific journals are arranged in order of decreasing productivity of articles on a given subject, they may be divided into a nucleus of periodicals more particularly devoted to the subject and several groups or zones containing the same number of articles as the nucleus, when the numbers of periodicals in the nucleus and succeeding zones will be as $1: n: n^2: n^3$, where n is a multiplier" (Bailón-Moreno, et al. 2005: 213). Simply stated, the law says that in a given subject field over a given period of time, (a) a few journals publish a relatively high per cent of articles in a field and (b) there are many journals that publish only a few articles each (Diodato 1994). Diodato (1994) states that a Bradford analysis may be used to test how well Bradford's law applies to a collection of items and sources (usually articles and journals) or to identify the core of journals in a field.

Yatsko (2012) observes that a number of authors have offered suggestions to (a) clarify the law (e.g. Brookes 1968; Bailón-Moreno, et al. 2005), (b) to offer its interpretation, (c) to analyse its conformity with the other laws and/or (d) to demonstrate the possibilities of its application for the analysis of various kinds of data (e.g. Drott, Mancall & Griffith 1979; Girap, Ashok & Bhanumurthy 2014). There are also those who have criticised the law, especially the formula, (e.g. Brookes 1985; Urquhart 1981; Heine 1998, all cited in Shenton & Hay-Gibson 2009). Criticism of this nature is not unexpected in scholarship. Despite the criticisms, Bradford's law remains a subject of much discussion, as witnessed in recent studies.

Most studies that have been conducted on Bradford's law have sought to determine the dispersion of articles in journals in order to identify core journals in a subject field or discipline (Andrés 2009; Hjørland & Nicolaisen 2005; Yatsko 2012). Recent studies that have been conducted to assess journal productivity in order to identify core journals in a subject field include those by Desai, Veras & Gosain (2018), Wahid & Idrees (2017), Neelamma & Gavisiddappa (2016), Yang, Tseng & Won (2016) and Singh & Bebi (2014). The law has also been applied to study the citations in a given journal in order to identify the list of core journals cited by authors publishing their works in the journal (Wahid & Idrees 2017). For example, Singh & Bebi (2014) used Bradford's law to study the cited references in PhD theses in social sciences at the University of Delhi and found that Bradford's law fitted the study. Lately, however, there have been attempts to study the application of the law beyond its original focus in assessing journal productivity. Tsay (2008) conducted a study by employing bibliometrics techniques and, more particularly, Bradford's law, to analyse the scattering and subject changes between the citing and cited literature in digital libraries. The author identified the core journals in the citing and cited literature in the subject domain and thereafter analysed the nature and subjects of the core journal in both categories. Among the conclusions drawn from the findings are that (a) only two journals in Bradford's nucleus were common in the two categories of journals, that is, citing and cited literature, and (b) while the citing core journals were devoted to the subject of the *application of computer and information technology to library implication*, the cited core journals addressed four main subject areas, namely *digital library orientation, general library and information science, new development in librarianship* and *library technology* (Tsay 2008: 713). It would have been interesting to determine if the aforementioned subject areas constituted the core subjects in the citing and cited literature, respectively.

Bradford's law has also been applied to examine how well the distribution of 'sitations' fits a Bradford-type distribution. In their study entitled "Sitation distributions and Bradford's law in a closed Web space", Faba-Pérez, Guerrero-Bote and Moya-Anegón (2003: 558) found that although the sitation distributions were consistent with those in previous experiments, the "plots of accumulated clusters of sitations [i.e. plotting the accumulated sitations against accumulated targets] and targets did not fit the typical Bradford distribution". Girap, Ashok & Bhanumurthy's (2014) study on the application of Bradford's law to the evaluation of book collection at the library of the Bhabha Atomic Research Centre is perhaps the closest to the current study. Using the Universal Decimal Classification class numbers, the authors identified 27 main subject headings to which a total of 94 450 books at the library belonged. The authors then ranked the subject headings according to the number of books in Bradford-type zones and noted that the subject headings formed three zones, with Bradford's nucleus (core zone) comprising two subject headings, the next zone consisting of five subject headings, and the last zone consisting of twenty (20) subject headings. When expressed according to the Bradford distribution, the pattern of the distribution of subject headings,

according to Girap, Ashok & Bhanumurthy's (2014) study, was 2:5:20, which was close to $1:2^1:2^2$, and in the Bradford-type distribution of $1: n^1: n^2$. Apparently, the dispersion pattern reported in the study by Girap, Ashok & Bhanumurthy (2014) does not fit Bradford's law perfectly, but it is close, as the third zone in their study consists of far more subject headings (i.e. 20) than the Bradford law suggests (i.e. 8).

In view of the above and despite the arguments that Bradford's law can be adopted to assess the distribution of literature according to subject terms (see Yatsko 2012), little research has been conducted in this area. The purpose of this study, therefore, is to investigate the possibility of applying Bradford's law of dispersion to determine core subject terms of information and communication technologies (ICTs) within the context of information and knowledge management (IKM) research. The specific objectives of the study are

- to determine the nature of dispersion of ICT-related subject terms and publications within the IKM context, using Bradford's law of scattering, from 1998-2017
- to identify the core subject terms in ICTs within IKM research from 1998-2017
- to determine the applicability of Bradford's law in the dispersion of the ICTs literature from 1998-2017

Methods and materials

We used EBSCO Discovery Service and EBSCOHost's Library, Information Science and Technology Abstracts (LISTA) and the Library and Information Science Source (LISS) databases to extract data. The former hosts many popular international databases and was, therefore, found most desirable. The latter two focus on LIS. An advanced search platform was used to conduct searches for articles published on ICTs using the search terms *information and communication technologies* and *information technologies* and variations thereof as well as *knowledge management*. The search was conducted within the subject fields in order to obtain results with high specificity and reliability. The search was limited to articles published in peer-reviewed and scholarly publications between 1998 and 2017.

The extracted data (i.e. subject terms) for each record were saved in text format to meet the requirements of the Bibexcel software, which was used to analyse the data in order to generate frequencies of term occurrence. As explained in the section above, the core subject terms were those that formed Bradford's nucleus. In order to determine the core subject terms, we followed the procedures outlined in Andrés (2009) and Singh & Bebi (2014) regarding the application of Bradford's law.

We ranked the subject terms and corresponding articles, as reflected in Appendixes A to D. Each of the tables in the appendixes consists of the number of subject terms, number of documents corresponding to the subject terms, cumulative subject terms, cumulative documents and the log (*ln*) of cumulative subjects. Secondly, we calculated the value of Bradford's constant (k) for each time period of study (i.e. 1998-2002, 2003-2007, 2008-2012 and 2013-2017) as follows:

$$\kappa = (e^{\gamma} \times \Upsilon_m)^{1/p}$$

where γ is Euler's number (i.e. $\gamma = 0.5772$), Υ_m is the maximum number of records for the highest ranked subject term, and p is the number of zones or Bradford's groups. Given the large number of subject terms, the number of zones was set at four (i.e. $p = 4$) for each year period. Consequently, Bradford's constant κ for each period of study was calculated as follows:

a) 1988-2002: $\kappa = (1.781 \times 283)^{1/4} = 4.738191$
b) 2003-2007: $\kappa = (1.781 \times 122)^{1/4} = 3.839334$
c) 2008-2012: $\kappa = (1.781 \times 59)^{1/4} = 3.201688$
d) 2013-2017: $\kappa = (1.781 \times 135)^{1/4} = 3.937762$

Thirdly, we calculated the number of subject terms that would constitute the core or Bradford's nucleus, using the following formula

$$r_0 = \frac{T(\kappa - 1)}{(\kappa^p - 1)}$$

where T is the total number of subject terms that are the subject of research in the ICT documents, κ is Bradford's constant and p is the number of zones or Bradford's groups. The value of the core subject terms (i.e. r_0), as obtained by using the above formula, for each time zone is provided in column three of table 1. The number of subject terms that would constitute the subsequent zones of Bradford's groups was calculated based on the number that constituted the core zone and Bradford's constant κ as follows:

$$r_1 = r_0 \times \kappa^1$$
$$r_2 = r_0 \times \kappa^2$$
$$r_3 = r_0 \times \kappa^3$$

This process of determining the number of subject terms that constituted each zone was repeated for each time zone. The results are presented in table 1. Once the number of the core subject terms was generated, we identified the subject terms from the ranked list of the terms in each time zone and plotted them in tables (see tables 2 to 5).

Results and discussion

Nature of dispersion of ICT-related subject terms and publications within the IKM context using Bradford's law of scattering

Table 1: Bradford zones for publications in ICTs research within the IKM context, 1998-2017

Period		Core (r_0)	Zone 1 (r_1)	Zone 2 (r_2)	Zone 3 (r_3)
1998-2002	No. of subject terms	7	34	160	759
	No. of articles	776	572	728	857
2003-2007	No. of subject terms	36	139	532	2043
	No. of articles	1380	1237	1592	2160
2008-2012	No. of subject terms	86	274	878	2811
	No. of articles	1902	1898	2255	2819
2013-2017	No. of subject terms	39	151	595	2344
	No. of articles	959	1295	1728	2509

Table 1 provides the results of the study according to Bradford's zones or groups. The table reveals that in 1998-2002, the core or nucleus consisted of seven subject terms, which were the subject of investigation or discussion in 776 articles, while the core in 2003-2007 comprised a total of 36 subject terms, which posted a total of 1 380 articles. In 2008-2012 and 2013-2017, there were a total of 86 and 39 subject terms, with some 1 902 and 959 articles in

Bradford's core or nucleus, respectively. The other zones, which can be termed "peripheral", produced a pattern that was similar to the core, but with a higher volume of subject terms and articles, as reflected in columns four to six. Table 1 further reveals that the number of subject terms that form the core in each time period varies, just as in the subsequent zones. The number of subject terms rose rapidly from 1998-2002 to 2008-2012, only for it to fall in 2013-2017. The least number of subject terms witnessed in the 2013-2017 time period can be attributed to retrospective indexing, where indexing time lag is a factor that influences the capturing of a record's metadata and/or bibliographic information in a database. The current study was conducted in April 2018, a situation that may have caused us to omit some of the articles that were published in the 2013-2017 time period, but were yet to be indexed in the two databases used as sources of data.

Core subject terms in ICTs research within IKM context, according to Bradford's nucleus

Having established the number of subject terms in each zone according to Bradford's law, we sought to identify and present the core subjects per the publication time period, as shown in tables 2 to 5. Table 2 presents the seven subject terms that comprised the core of Bradford's distribution for the period 1998 to 2002. Leading the pack is *technology – information services,* with 283 (24.76%) articles, followed closely by *information technology*, which yielded 263 (23.01%) articles. It is evident that the core areas of research during this period were associated with the application of information technology (as opposed to ICTs in broad terms) in the provision of information services. The occurrence of *information society* among the core subject terms testifies to the hype around the concept "information society" in the 1990s (Technopolitics Working Group 2015). It should be acknowledged that one of the characteristics of an information society is the presence and application of information technology as an enabler of accessing, processing or organising, storing, disseminating or transferring and manipulating information (Balan 2013). The presence of *electronic publications* among the seven core subject terms is not surprising, as e-publications are products of ICTs. The *internet* seems to have been the main enabler for library services during the 1998-2002 period of study.

Table 2: Core subject terms in ICT research, 1998-2002 (N = 1143)

	SUBJECT TERM	f	%
1	Technology – Information services	283	24.76
2	Information technology	263	23.01
3	Information society	58	5.07
4	Internet	53	4.64
5	Libraries	50	4.37
6	Library science	39	3.41
7	Electronic publications	30	2.62

The 2003-2007 period saw the number of core subject terms not only increase to 36, but also the falling away of some of the core terms that existed in Bradford's core in 1998-2002. All but one subject term in table 2 were excluded from the list of core subject terms in 2003-2007, implying that out of the 36 subject terms that constituted Bradford's core in 2003-2007, 35 were new. If we exclude terms that do not necessarily explain the application of ICTs in the context of IKM (e.g. *conferences & conventions, developing countries, prefaces & forwards* and *case studies*), the core subject terms in 2003-2007 are largely associated with business/trade/commerce and the management of organisations. We further note that no specific ICTs are mentioned among the core subject terms. This may imply that all articles on

ICTs were indexed under the headings *technology* or *information technology*, or that no ICT-specific subject term was core in the period of study. Nevertheless, we note that *management information systems*, defined as a "set of procedures that collects (or retrieves), processes, stores, and disseminates information to support decision making and control" (Laudon & Laudo 2003) is the only ICT-oriented term that appears in the core during the 2003-2007 period. Its presence explains the occurrence of the other terms, which are largely associated with the management of and decision-making processes in organisations and public institutions. Another system-oriented term that appears in table 3 is *instructional systems*, which is closely associated with education and training and such subject terms as *higher education, teaching, educational innovations* and *teacher training*. This implies a new set of subject terms, or a cluster of subject terms, that is associated with and can be used to explain the application of ICTs within the IKM sub-field.

Table 3: Core subject terms in ICT research, 2003-2007 (N = 2813)

SUBJECT TERM	f	%	SUBJECT TERM	f	%
Technology	122	4.34	Decision making	29	1.03
High technology	106	3.77	Enterprise resource planning	28	1.00
Conferences & conventions	95	3.38	Public institutions	27	0.96
Management information systems	83	2.95	Higher education	26	0.92
Business enterprises	75	2.67	Teaching	26	0.92
Management	72	2.56	Business	26	0.92
Electronic commerce	55	1.96	Prefaces & forewords	25	0.89
Medical care	40	1.42	Educational innovations	24	0.85
Industrial management	39	1.39	Work environment	23	0.82
Developing countries	37	1.32	Organisational learning	22	0.78
Intellectual capital	33	1.17	Management science	21	0.75
Contracting out	32	1.14	Teacher training	20	0.71
Public administration	32	1.14	Economic development	20	0.71
Information technology	31	1.10	Personnel management	19	0.68
Business planning	31	1.10	Instructional systems	19	0.68
Corporate culture	30	1.07	Investments	18	0.64
Globalisation	29	1.03	Case studies	18	0.64
Organisation	29	1.03	College teachers	18	0.64

Table 4 consists of 86 subject terms in the Bradford nucleus for the period 2008-2012. As in the previous time period of 2003-2007, the period represented in table 4 witnessed some new additions, while other subject terms fell away. Out of the 36 subject terms that formed the core in 2003-2007, six did not feature in 2008-2012, namely *organisation, enterprise resource planning, public institutions, business, prefaces & forewords* and *management science*. The rest of the subject terms, numbering 29, featured in Bradford's core for 2008-2012, as provided in table 4. Therefore, the implication is that, of the 86 core subject terms in table 4, 57 were new. Unlike in the previous years, no single subject term was clearly dominant in 2008-2012. The top-ranking terms yielded articles that were close in terms of their (i.e. subject terms) frequency of occurrence in the databases. The period represented in table 4 saw the dominance of issues revolving around management science, trade, education and medicine.

Table 4: Core subject terms in ICT research, 2008-2012 (N = 3265)

SUBJECT TERM	f	%	SUBJECT TERM	f	%
Conferences & conventions	59	1.81	Attitude (psychology)	18	0.55
Developing countries	58	1.78	Small business	17	0.52
Management information systems	52	1.59	Interpersonal relations	17	0.52
Decision making	49	1.50	Work environment	17	0.52
Online social networks	48	1.47	Organisational behaviour	17	0.52
Electronic commerce	44	1.35	Diffusion of innovations	17	0.52
Business enterprises	42	1.29	Social interaction	17	0.52
Social networks	42	1.29	Risk management in business	16	0.49
Industrial management	39	1.19	Communication in learning & scholarship	16	0.49
Medical care	38	1.16	Customer services	16	0.49
Management science	35	1.07	Investments	16	0.49
High technology	35	1.07	Economic development	16	0.49
Business planning	35	1.07	Innovation management	16	0.49
Higher education	34	1.04	Government policy	15	0.46
Corporate culture	32	0.98	Standardisation	15	0.46
Technology	31	0.95	Stakeholders	15	0.46
Empirical research	29	0.89	User-centred system design	15	0.46
Case studies	29	0.89	Leadership	15	0.46
Data security	28	0.86	Computer engineering	14	0.43
Management	27	0.83	College students	14	0.43
Organisational learning	27	0.83	Medical communication	14	0.43
Theory of knowledge	26	0.80	Instructional systems	14	0.43
Social sciences	25	0.77	Cellphones	14	0.43
Intellectual capital	25	0.77	Teaching	14	0.43
Educational innovations	25	0.77	Culture	13	0.40
College teachers	25	0.77	Human capital	13	0.40
Information technology	24	0.74	Health education	13	0.40
Public administration	24	0.74	Collaborative learning	13	0.40
Books – reviews	23	0.70	University faculty	13	0.40
Technology & society	22	0.67	Career development	13	0.40
Project management	22	0.67	Medical technology	12	0.37
Teaching methods	22	0.67	Research institutes	12	0.37
Technology acceptance model	21	0.64	Standards	12	0.37
Contracting out	21	0.64	European Union	12	0.37
Public sector	21	0.64	Interviewing	12	0.37
Organisational change	21	0.64	Competitive advantage in business	12	0.37
Personnel management	20	0.61	Local government	12	0.37
Information technology conferences	20	0.61	Regression analysis	12	0.37
Knowledge management research	19	0.58	Communication in organisations	12	0.37
Business communication	19	0.58	Teachers	12	0.37
Quality of service	19	0.58	Municipal services	12	0.37

Methodology	18	0.55	Academic libraries – Nigeria	12	0.37
Globalisation	18	0.55	Educational programmes	12	0.37

The top-ranking subject terms during the 2008-2012 time period were similar to those occurring in the previous time period, despite the slight differences in terms of their frequencies of occurrence. *Management information systems, electronic commerce, business enterprises, industrial management* and *organisational learning* were among the top-ranking terms that occurred in the two core zones of 2003-2007 and 2008-2012. Information technology-based terms popped up in the latter period, with *management information systems, online social networks, social networks, high technology, technology, information technology, technology & society* and *user-centred system design* appearing frequently in the ICT literature. Furthermore, the occurrence of *technology acceptance model* and *diffusion of innovations* may reveal the studies' orientation, namely the acceptance and diffusion of ICTs in management, education, trade, organisational learning, knowledge management and medical care, among others.

Table 5: Core subject terms in ICT research, 2013-2017 (N=2258)

SUBJECT TERM	f	%	SUBJECT TERM	f	%
Information technology in medicine	135	5.98	Organisational change	19	0.84
Decision making	45	1.99	Knowledge management research	19	0.84
Diffusion of innovations	43	1.90	Online social networks	19	0.84
Higher education	41	1.82	Work environment	18	0.80
Information technology industry	32	1.42	Technology	18	0.80
Organisational performance	30	1.33	Interviewing	17	0.75
Medical care	30	1.33	Data analysis software	17	0.75
Telemedicine	26	1.15	Comparative studies	17	0.75
Business enterprises	26	1.15	Self-efficacy	17	0.75
Structural equation modelling	25	1.11	Industrial management	17	0.75
Financing of research	23	1.02	Theory of knowledge	16	0.71
Public sector	23	1.02	Learning strategies	16	0.71
Public administration	22	0.97	Information technology periodicals	16	0.71
Mathematical models	21	0.93	Cellphones	15	0.66
Project management	21	0.93	Data security	15	0.66
Electronic commerce	21	0.93	Theory	15	0.66
Social networks	21	0.93	Medical communication	15	0.66
Competitive advantage in business	20	0.89	Small business	15	0.66
Attitude (psychology)	20	0.89	Economic development	14	0.62
Personnel management	19	0.84			

Bradford's core concepts associated with ICTs in 2013-2017 totalled 39, as shown in table 5. One-third (13) of the subject terms were common in the two time periods of 2008-2012 and 2013-2017.

Does the disperson of ICTs subject terms in IKM research fit Bradford's law?

In this section, we focus our attention on two aspects to determine whether the data fit Bradford's law so that we can gauge the suitability of this law in the identification of the core subject terms on ICTs as indexed in the LISTA and LISS databases. In the first instance, we compute the Bradford multiplier κ in each zone by dividing the number of subject terms in the subsequent zone by the number in the previous zone, i.e.

$$\kappa = \frac{r_n}{r_{n-1}}, \text{ the numerator excludes the nucleus value}$$

If the multiplier is similar to Bradford's constant, as provided in the methodology section, then the data are said to fit Bradford's law. A comparison between the Bradford constant (i.e. κ) in table 6 and Bradford's constant (i.e. κ) as shown in the methods and materials section above reveals that the two values are similar across all the zones. For instance, whereas the constant for the time zone 1998-2002 is 4.738191, the constant in the same time zone in table 6 ranges from 4.7 to 4.9. When the values are rounded off to the nearest whole figure, they all equal 5.0, therefore registering a perfect semblance.

Table 6: Bradford's constant (κ) for each Bradford's zone per year period

Period		Core (r_0)	Zone 1 (r_1)	Zone 2 (r_2)	Zone 3 (r_3)
1998-2002	No. of subjects	7	34	160	759
	κ	-	4.8571	4.7059	4.7438
2003-2007	No. of subjects	36	139	532	2043
	κ	-	3.8611	3.8273	3.8402
2008-2012	No. of subjects	86	274	878	2811
	κ	-	3.1860	3.2044	3.2016
2013-2017	No. of subjects	39	151	595	2344
	κ	-	3.8718	3.9404	3.9395

Table 7: Zones and proportions of ICTs subject terms, 1998-2017

Period		Core (r_0)	Zone 1 (r_1)	Zone 2 (r_2)	Zone 3 (r_3)	Bradford's multiplier (n)
1998-2002	No. of subjects	7	34	160	759	
	Proportion	1	4.86	22.86	108.43	5
2003-2007	No. of subjects	36	139	532	2043	
	Proportion	1	3.86	14.78	56.75	4
2008-2012	No. of subjects	86	274	878	2811	
	Proportion	1	3.19	10.21	32.69	3
2013-2017	No. of subjects	39	151	595	2344	
	Proportion	1	3.87	15.26	60.10	4

In the second instance, we determine the ratio of the number of subject terms in one Bradford zone to the value in the subsequent zone to determine whether the proportion of the data in this study fits Bradford's pattern, that is, $1: n: n^2: n^3$ ... Table 7 provides the proportional distribution of the values, expressed as ratios in the pattern of Bradford's expression of the law.

The proportional pattern of the distribution of the subject terms for each time period is therefore as follows:

a) **1998-2002:** 1:4.86:22.86:108.43 or 1:5:23:108, which is close to and can be expressed as $1: 5^1: 5^2: 5^3$. The multiplier (n) for the period 1988-2002 is 5.

b) **2003-**2007: 1:3.86:14:78:56.75 or 1:4:15:57, which is close to and can be expressed as $1: 4^1: 4^2: 4^3$. The multiplier (n) for the period 2003-2007 is 4.

c) **2008-2012**: 1:3.19:10.21:32.69 or 1:3:10:33, which is close to and can be expressed as $1: 3^1: 3^2: 3^3$. The multiplier (n) for the period 2008-2012 is 3.

d) **2013-2017**: 1:3.87:15.26:60.10 or 1:4:15:60, which is close to and can be expressed as $1: 4^1: 4^2: 4^3$. The multiplier (n) for the period 2013-2017 is 4.

Conclusion

The application of Bradford's law beyond its originally intended purpose has excited a few researchers in recent times. We were able to find two studies that have attempted to assess the applicability of the law beyond journals as the units of analysis despite some scholars' observations that the law can be used in diverse ways. This study sought to explore the use of Bradford's law to determine the core subject terms in ICTs research within the IKM subject domain. A total of seven subject terms were considered as the core in the period 1998 to 2002, while the period 2003-2007 yielded 36 subject terms in the core zone. The number of core subject terms for the next two time zones of 2007-2012 and 2013-2017 were 86 and 39, respectively. The subject terms that formed each core, from 1998-2017, differed both in number and constitution. In all the cases after the 1998-2017 time period, most subject terms that appeared in subsequent core zones were new. Whether this pattern is normal is a subject for further research. However, the shifting of the subject terms that constituted the core in each case (see tables 2-5) may be an indication of an evolving subject domain, change of research focus, the interdisciplinary nature of the research fields of ICTs and KM, or dynamic indexing services. As the ICTs remain relevant and important tools for IKM, among other aspects affecting human life, such as health, the pattern of distribution of subject terms in the core zones is likely to persist in research. New concepts will continue emerging, while others will move from the core to the peripheral zones.

The proportional distribution of subject terms reveals that the number of subject terms in each Bradford group is proportional to $1: n: n^2: n^3$. Tables 6 and 7 provide Bradford's constant and multiplier in each time period, respectively. We aver that the data in this study fit Bradford's law and, therefore, the law can be used not only to determine the number of ICTs subject terms that are core IKM research, but also to identify the core subject terms in a specific research area. However, it would be interesting to conduct the same study using different subject domains as case studies to assess the applicability of the law. Such an exploration would provide adequate evidence to make informed conclusions about the application of the law in assessing other units besides the dispersion of the literature in journals.

Having confirmed that the data fits the Bradford's proportional distribution, we hasten to add that the distribution of articles does not entirely correspond to Bradford's assertion that groups or zones into which the articles are dispersed would *contain the same number of articles as the nucleus*. Table 1 reveals that the core consisted much fewer articles than the rest of the zones except for the first zone, which, in three out of the four time-periods of study, contained fewer articles than the core. The difference between zone with the largest

number of articles and those in the core (or nucleus) are as follows: 1998-2002 (81), 2003-2007 (780), 2008-2012 (917), and 2013-2017 (1550). This pattern can be attributed to the fact that unlike in the case of distribution of articles in journals where a given article belong to one journal, in the subject analysis presented in this study, articles may belong to more than one subject term; that is, an article can be indexed in or using several subject terms. A mathematical formula or model to explain such a phenomenon may be challenging to formulate. Further research is therefore recommended to investigate this phenomenon as well as the shape of the Bradford's curve in subject analysis in a field. Similar studies, testing the applicability of Bradford's law in other fields other than IKM, are recommend for purposes of comparing the validating the results of the current study.

Implications of the study for information practice

The implications of the application of Bradford's law in information practice are well documented. Qiu et al (2017) have argued that the law has both theoretical and practical applications. The law's application in practice is diverse and includes the determination of core journals, literature searches, investigation of monograph distribution, maintenance of dynamic collections, measurement of the integrity of search tools, guiding users to use journals, guiding the work of journal subscription (Qiu et al. 2017) and as a tool in developing information systems (Von Ungern-Sternberg 2000).

Thus, the current study has demonstrated a possible area and manner in which Bradford's law can be applied in practice beyond its original context of application. Broadly, the findings of the current study can be used in the development of an information system, as suggested by Von Ungern-Sternberg (2000), as well as in literature research and collection development, as noted by Qiu et al. (2017). Furthermore, we note that Bradford's law can inform selection and di-selection of items besides journals; determine the growth and development of science and knowledge by, for example, their auditing, mapping and forecasting; provide a tool for data analysis (e.g. subject and keyword analysis); support information retrieval and knowledge organization; determine disciplinary boundary crossing; and enable the determination of databases and search engines, among others. The aforementioned areas of the Bradford law's application are central in information/knowledge processes, right from information/knowledge creation to their use and re-use.

In terms of the discipline/field-specific relevance of the current study, the importance of KM in enhancing organizational performance cannot be overemphasized. It has been observed that KM is a young but growing and dynamic multidisciplinary field (Ebrahim, Fetrati & Pezeshkan 2016). As a result, scholars from diverse fields and disciplines have converged to study or investigate various aspects of KM. The role and application of ICTs in KM is one of such issues that have become a central theme in KM research (van der Velden 2002; Kör 2017). The field can benefit from this study in terms of, for example, curriculum development, thesaurus construction and subject organisation and description in the area of ICTs as they relate to KM.

References

Andrés, A. (2009). *Measuring academic research: how to undertake a bibliometric study*. Oxford: Chandos.

Bailón-Moreno, R., Jurado-Alameda, E., Ruiz-Banós, R. & Courtial, J.P. (2005). Bibliometric laws: empirical flaws of fit. *Scientometrics,* 63(2): 209-229.

Balan, S-G. (2013). Information society: context, manifestations, socioeconomic implications. *Internal Auditing & Risk Management*, 8(2): 1-11.

Borgatti, S. P. & Everett, M. G. (1999). Models of core/periphery structures. *Social Networks*, 21: 375-395.

Brookes, B.C. (1968). The derivation and application of the Bradford-Zipf distribution. *Journal of Documentation*, 24(4): 247-265.

Desai, N., Veras, L. & Gosain, A. (2018). Using Bradford's law of scattering to identify the core journals of pediatric surgery. *Journal of Surgical Research*, 229: 90-95

Diodato, V. (1994). *Dictionary of bibliometrics*. London: Routledge.

Ebrahim, N. A., Fetrati, M. A., & Pezeshkan, A. (2016). Major trends in knowledge management research: a bibliometric study. *Scientometrics,* 107(3): 1249-1264.

Faba-Pérez, C., Guerrero-Bote, V. P. & Moya-Anegón, F. (2003). "Sitation" distributions and Bradford's law in a closed Web space. *Journal of Documentation*, 59(5): 558-580, https://doi.org/10.1108/00220410310499582.

Girap, P., Ashok, T. & Bhanumurthy, K. (2014). Application of Bradford's law to the evaluation of book collection of library of Bhabha Atomic Research Centre. *Journal Scientometric Research,* 3(1): 22-27.

Kör, B. (2017). Revealing trends in knowledge management: from 2010 to 2015. *Instanbul University Journal of the School of Business*, 46: 18-30.

Neelamma, G. & Gavisiddappa, A. (2016). Examination of Braford's law in the field of zoology literature: a citation study from 2005-2014. *Journal of Advancements in Library Science*, 3(3): 32-46.

Ocholla, D. N., Onyancha, O. B. & Britz, J. (2010). Can information ethics be conceptualized by using the core/periphery model? *Journal of Informetrics*, 4: 492-502.

Onyancha, O. B. & Mokwatlo, K. (2012). How does LIS perceive records management? A trend and core/periphery model analyses. *ESARBICA*, 31: 111-124.

Onyancha, O. B. & Ocholla, D. N. (2009). Conceptualising 'knowledge management' in the context of library and information science using the core/periphery model. *South African Journal of Information Management,* 11(4).

Qiu, J., Zhao, R., Yang, S. & Dong, K. (2017). *Informetrics: theory, methods and applications.* Singapore: Springer.

Shenton, A. K. & Hay-Gibson, N. V. (2009). Bradford's law and its relevance to researchers. *Education for Information*, 27: 217-230.

Singh, K.P. & Bebi, D. (2014). Application of Bradford's law on journal citations: a study of PhD theses in social sciences of University of Delhi. *Annals of Library and Information Studies*, 61: 112-120.

Technopolitics Working Group. (2015). Tracing information society – a timeline. *Technopolitics*, 4(4): 50-53.

Tsay, M-Y. (2008). Subject change between citing and cited literature on digital libraries. *The Electronic Library*, 26(5): 702-715, https://doi.org/10.1108/02640470810910729.

Van der Velden, M. (2002). Knowledge facts, knowledge fiction: the role of ICTs in knowledge management for development. *Journal of International Development*, 14(1): 25-37.

Von Ungern-Sternberg, S. (2000). Bradford's law in the context of information provision. *Scientometrics*, 49(1): 161-186.

Wahid, N. & Idrees, H. (2017). Application of Bradford's law on Pakistan Journal of Psychological Research. *Pakistan Journal of Psychological Research*, 32(1): 1-17.

Yang, J.M., Tseng, S.F. & Won, Y.L. (2016). A bibliometric analysis on data mining using Bradfrd's law. In: Juang, J. (ed). *Proceedings of the 3rd International Conference on Intelligent Technologies and engineering Systems (ICITES2014). Lecture Notes in Electrical Engineering,* 345. Cham: Springer.

Yatsko, V. A. (2012). The interpretation of Bradford's Law in terms of geometric progression. *Automatic Documentation and Mathematical Linguistics,* 46(2): 112-117.

NETSCITY: a geospatial application to analyse and map world scale production and collaboration data between cities

Marion Maisonobe[1], Laurent Jégou[2], Nikita Yakimovich[2], and Guillaume Cabanac[3]

[1] *marion.maisonobe@cnrs.fr*
CNRS, Géographie-cités UMR 8504 CNRS, Paris (France)

[2] *laurent.jegou@univ-tlse2.fr*, *nikita110598@mail.ru*
University of Toulouse, LISST UMR 5193 CNRS, Université Toulouse Jean Jaurès, Toulouse (France)

[3] *guillaume.cabanac@univ-tlse3.fr*
University of Toulouse, Computer Science department, IRIT UMR 5505 CNRS, Toulouse (France)

Abstract

We present NETSCITY, an online application to analyse and visualise world scale scientific production and collaboration data between cities. Contrary to existing tools that mainly focus on displaying co-occurrence networks, NETSCITY especially focuses on processing the geographical information comprised in bibliometric data. NETSCITY proposes a fully integrated solution to parse and clean the authors' addresses, comprised in a set of references, geocoding them at the city level, clustering them at the requested level of analysis (urban areas or countries) and mapping them either on a world base map or in a relational space. In the first part of the paper, we stress the originality and design of the NETSCITY application in terms of geocoding, clustering, counting, and mapping methods. In the second part, we detail its main functionality as a geoweb application. Eventually, we show the results one can obtain by applying NETSCITY on a set of references extracted from the online version of the Web of Science Core Collection.

Introduction

Practitioners in the field of bibliometrics contributed a variety of software tools to simplify the processing and mapping of bibliographic data. We can think of the online dashboards available through the Web of Science Core Collection and Scopus websites. There is also a variety of free bibliometric mapping software tools such as VOSViewer (van Eck & Waltman, 2010) and CiteSpace (Chen, 2006). According to the review proposed by Cobo et al. (2011), which focuses on these latter tools together with Bibexcel, CoPalRed, IN-SPIRE, Leydesdorff's programs, Network Workbench Tool, Sci^2 Tool, and Vantage Point, existing software tools offer four different types of bibliometric analyses: burst detection, geospatial analysis, network analysis, and temporal analysis. Among the nine tools reviewed by Cobo et al., network analysis is the most prominent one. Indeed, existing tools primarily allow computing and/or mapping of bibliometric networks of words, authors, and journals derived from co-occurrence data. Geospatial analysis, on the contrary, is the less common type of available analyses. It is included in only three of the reviewed software tools: CiteSpace, Sci^2 Tool, and VantagePoint. The first two allow network data visualisation over a world map using Google Earth Maps or Yahoo! Maps. They also include geocoding capabilities but the choice of the spatial resolution is only dual: either the data are clustered at the country level, or they are clustered at the street level – street addresses used in the publications are then converted to coordinates in latitude and longitude (Chen, 2016). In addition to geocoding and mapping capabilities, NETSCITY, the online application for geographic analysis of research output that we present in this article, features an intermediary level of geographic aggregation, which is the urban area level. Actually, NETSCITY tackles four main issues in the business of mapping bibliographic data:

the geocoding accuracy, the geographic aggregation, the counting/fractioning issue, and the mapping issue.

NETSCITY is a free geoweb application developed to promote the spatial scientometrics method designed by our team to research on the world geography of scientific production and collaboration. Over the past 10 years, our team has been working on geocoding, clustering, and mapping the contemporary geography of scientific activities using bibliometric data. To measure production share and collaboration intensity, we delineated urban area perimeters covering geocoded addresses. This aggregation step proved necessary to work with statistically comparable geographic entities at the world level, which administrative municipalities or street addresses are not. While sharing a set of 495 urban area perimeters in an open access publication (Maisonobe et al., 2018) we have developed NETSCITY, which allows any user:

- to upload a set of references to scientific materials;
- to geocode them;
- to aggregate them at the urban area and the country levels;
- to compute the number of publications per spatial unit and the number of collaborations between spatial units using various counting methods;
- to visualise these variables both on a geographic and on a network map.

At each step, the user is free to download the results of the current process. Some users can be interested in the results of the geocoding process only, others in the results of the aggregation process, others in the results of the counting process, and others in the results of the mapping process. NETSCITY thus fulfils a various number of spatial analytics purposes. We designed it is as a key application to ensure a better reproducibility in the spatial bibliometric field (Frenken et al., 2009; Cobo et al., 2018). We also plan to release the source code of the application.

In this paper, we first discuss the inputs of NETSCITY for the spatial bibliometric field. Then, we describe the inner workings of NETSCITY: from upstream geocoding down to mapping processes. Next, we show the results of NETSCITY when applied to a specific set of references. We conclude with future works involving additional levels of geographic aggregation, counting methods, and mapping options.

Mapping the geography of science

Geocoding

Mapping scientific activities is a five-century-old craft, with seminal maps of science showcased in Börner's (2010) *Atlas of Science*. In 2010, Loet Leydesdorff and Olle Persson (2010) published a landmark article discussing the opportunities of geocoding and mapping Web of Science and Scopus data at the city and institution levels. One year before, Frenken et al. (2009), were identifying a spatial turn in scientometrics that has been confirmed later (Frenken & Hoekman, 2014). The improving capabilities and availability of online geocoding tools, the trend toward territorialised policies of science, and the renewed enthusiasm for global benchmarking of cities and higher education institutions prompted this spatial turn. At the same time, our research team, mainly based in Toulouse (southwestern France), identified a need for strengthening the ties between bibliometrics and geomatics. Processing the spatial information included in bibliometric datasets by relying only on the results of online geocoding tools proved insufficient, such as Google Maps API (Jégou, 2014). Many errors result from trying to geocode the entire address strings included in bibliometric datasets. Let us consider the following example: an author of a 2009 publication reported the following address: "Chang Gung Mem Hosp, Tao Yuan, Taiwan". It corresponds to the Chang Gung Hospital, in Taoyuan District, Taoyuan county of Taiwan Island, 15 km west of the capital, Taipei. Google Maps API hesitates between the hospital itself and the Chang Gung Memorial metro station. Other web geocoding

services confuse this address with several locations in the district. Indeed, the district name is similar to the county name and to the island name.

If the required level of aggregation is the urban level, we proved that it is more relevant to begin by spotting the portion of the address string corresponding to the city, the province / state and the country. The latter is the easiest to find since it always ends the string. As shown in the aforementioned example, tagging the city and the province is a more complicated task because their order of appearance might differ from one address to another and the province is not always specified (non-federal countries such as France do not use them in postal addresses). Even by pre-processing the data to ensure better geocoding results, some errors or missing values still occur: some of them result from homonyms, which require additional spatial information, other result from misspelled toponyms. Here we should highlight that geocoding as we performed it is a semi-automatic process: a geographer expert reviews the results, focussing on the records affiliated to the most prominent scientific cities or the most prone to potential errors (we devoted additional work on addresses having undergone an alphabet transliteration and/or with postal system specificities like Indonesia or Taiwan). After years of geocoding work on the entire contents of the Web of Science Core Collection and on Scopus datasets, our research team improved the geocoding results one obtains when using these databases (Jégou, 2014). Capitalizing on this previous work, the geocoding tool implemented in NETSCITY detects and corrects much of the common misspellings found in Web of Science and Scopus data (by simplifying the given addresses, comparing them with a list of known misspellings variants and, in the absence of a match, by using internal and external online geocoding services) — these stem from authors, publishers, typists working for such database vendors. For the remaining errors, NETSCITY users can amend the records as they see fit (see next section).

Geographic levels of analysis

To study the world geography of science, our team strived to determine an adequate urban level of data aggregation. For studying the geography of science before 2010, most scholars were relying on the country level and only a few studies tackled it at the urban level. The latter has attracted a growing interest since the 2000s following city and regions' empowerment and the multiplication of city rankings (Bornmann & Moya-Anegón, 2018). Two types of studies can be distinguished: those focusing on a limited number of urban regions (e.g., Matthiessen et al., 2010; Hoekman et al., 2010; Nomaler et al., 2014); and those encompassing the publishing localities of the entire world (e.g., Waltman et al., 2011; Pan et al., 2012; Csomós, 2018). In the first case, the authors tend to reuse existing sets of administrative perimeters: United States MSA, European NUTS and, in the second case, the authors do not tend to aggregate the geocoding localities. To get a global panorama of the world scientific production, neither of these two approaches is satisfying. Existing sets of urban area perimeters fulfil specific purposes (whether they are defined for a national or a continental scope, or they are limited to the most populated areas only). Combining existing sets in order to encompass the entire world production might be a solution but it leads to comparative biases since the definitions used to produce these various sets are diverse. Not aggregating the data at all also leads to comparative biases since municipalities' boundaries depend on very heterogeneous criteria differing from one country to another (Maisonobe et al., 2018). For example, not aggregating the data will lead to count apart the numerous publications authored from Bethesda (Maryland, USA) of those of Washington City (DC, USA). Bethesda is a suburb of Washington, which include important medical institutions such as the NIH.

To overcome these issues our team designed a unique set of urban area perimeters encompassing all publishing localities identified after geocoding the entire 1999-2014 contents

of the Web of Science Core Collection. Our urban delineations consider the distribution of the world population density and the Euclidean distance between publishing localities. So far, NETSCITY allows its users to aggregate their data at the level of this set of urban area perimeters as well as at the country level.

Counting method

The next issue addressed by NETSCITY is the counting conundrum. Since about one third of the entire world publications are authored from more than one urban area, one needs to decide how to measure urban areas' contribution to science production. There exist many ways of assigning a number of publications to statistical units. Gauffriau et al. (2008) gives a rich overview of existing methods and highlight the need for more transparency on the methods used since the choice of a counting method significantly influences research findings in bibliometrics. When focusing on the geography of science, the elementary unit of analysis can be the author or the institution, but given the biases of existing sources, it is more often the address. Indeed, in pre-2008 bibliometric records, the list of addresses is not always linked to the list of authors. As a result, it is possible to derive a number of authors per publication as well as a number of addresses but not a number of authors per address per publication. In addition, authors' addresses do not necessarily correspond to authors' institutions since an address may refer to several institutions. Conversely, different addresses may refer to the same institution with varying research teams or university departments. Unless pre-processing the data accordingly, the postal address is the most elementary counting unit of a bibliometric dataset for geographic purpose. As a result, there are two main ways of measuring the scientific production of an urban area: adding up the number of addresses per urban area involved in the publication or considering the number of different urban areas involved in the publication. Similarly, to measure the scientific production of a country, we can consider three different values: the number of addresses per country involved in the publication, the number of urban areas per country involved in the publication, or the number of involved countries. In some disciplines such as in chemistry, only the first address or the corresponding address might be relevant. Other addresses would not be counted in this latter case. In addition to the choice of a counting unit, the counting issue requires to arbitrate between full and fractional countings (Van Hooydonk, 1997). In the first case, we consider the total number of addresses/urban areas/countries while in the second case, we fraction the credit of the publication (i.e., one unit) so that the sum of each fractioned credit total one. Since fractioning avoids counting a single country contribution multiple times, it is the most preferred method in bibliometrics. As shown by Leydesdorff & Park (2017), the choice of a counting method applies both to production data (computing a number of publications per spatial entity) and to collaboration data (computing a collaboration weight between co-authoring entities). With NETSCITY, it is possible to control for the computation of both indicators.

Mapping issue

Mapping issues addressed in bibliometric research are traditionally limited to the visualisation of co-occurrence networks in a relational space. Waltman & van Eck (2010) distinguish two main types of bibliometric maps, namely graph-based maps and distance-based maps. The former refer to mapping the presence or absence of a relation between entities while the latter refer to mapping the distance between entities according to the intensity of their relations. There exists a large range of similarity metrics to compute a relational distance between bibliometric entities: cosine, association strength, Jaccard index, Pearson index and, of course, the raw co-occurrence number (either full or fractional as shown in the previous sub-section) (Boyack et al., 2005; van Eck & Waltman, 2009). When used to produce a distance-based map, similarity metrics can be selected according to different mapping algorithms. Kamada Kawai and

Fruchterman Reingold are the most commonly used techniques in the field of bibliometric as well as in the broader field of network analysis. To adapt to power-law characteristics of most bibliometric distributions, Waltman et al. (2010) propose a unified way of mapping and clustering bibliometric networks. VoSViewer as well as other network analysis software such as Pajek include this method. Compared to this latter issue, mapping issues related to the production of geographic maps are mostly overlooked. The world maps used in spatial bibliometrics are often screenshots taken from Google Earth or Yahoo! Maps' base maps. Instead of benefitting from the capabilities of thematic cartography to map production and collaboration data, there seems to be a preference for online and interactive visualisation. However, we believe the two types of graphic representation – static and interactive – should be valued and different visualisation methods must be chosen accordingly, within the context of geomatics. In this respect, the literature in GIS and cartography about flow maps and the development of geoweb applications might be of relevance (Dodge et al., 2011). While NETSCITY displays interactive geographic and network maps, it also allows the downloading of all underlying data necessary to generate static maps. This feature of the application is about to include more adjustable parameters, interactive linked graphs, and a vector graphics output. In forthcoming developments, NETSCITY will also give access to the Netmap visualisation interface[1] allowing the simultaneous exploration of the geographic and the network maps: by clicking on an item on the graph, the item and its relations are highlighted on the map and vice versa (Maisonobe & Jégou, 2018). As demonstrated in Andurand et al. (2015) and Bach et al. (2015), combining several types of graphic representation enriches the data exploration and analysis experience.

In what follows, we explain how to use the NETSCITY geoweb application and we showcase results obtained by applying NETSCITY to a set of references extracted from the Web of Science.

Features implemented in NETSCITY

In this section, we provide a description of the geoweb application from the data import and processing step to the visualization step. We also discuss technical implementations and directions for improvements throughout the section.

NETSCITY is designed for different information needs and types of users:

- The researchers interested in mapping their own field of research or their own publication record.
- The librarians interested in organizing sets of references and mapping the geographic relations between scientific references.
- The policy officers or international experts interested in scientific evaluation.
- The students or journalists interested in mapping a field of knowledge or the production of a specific institution, author, or place.
- The science studies researchers interested — as we are — in mapping the geography of science.

The home page of NETCITY, presenting the project, is available following this web link: https://irit.fr/netscity

The main hub features four types of actions are available: 1) import a set of references extracted from the Web of Science or Scopus online websites, or personal CSV records, 2) explore the results (lists or data-visualizations) 3) export the results 4) visit the help pages.

[1] Source code of the prototype: https://gitlab.com/ljegou/netmap
Demo: http://www.geotests.net/test/net_map/science_mep.html

Importing a dataset from Web of Science, Scopus or else

In the current version, users can only import one type of Web of Science file. From the Web of Science web interface, they must download a set of references using the following method: 1) Search for a list of records to export and select "Save to Other File Format", 2) Export all records returned (up to 500) by entering "1" to "[the total number of records]", 3) Select "Full Record and Cited References" to include addresses and times cited information, 4) Select "Tab-Delimitated (Mac/Win, UTF-8)" from the "save" drop-down menu. Similarly, for a Scopus file, the procedure is the following: 1) Click on the Export icon[2], 2) From the pop-up menu, select "CSV" format for your file format and "Citation" as well as "Bibliographical information" as the information to export, 3) Click "Export", select "CSV - Only the first 2,000 documents" and click "Export" again.

It is also possible to import a personal CSV file. To do so, users must specify the name of the ID column and the Year column. The name of the columns storing the spatial information is also needed. Several options are possible since the spatial information can be stored:
- In entire postal addresses' strings,
- In three pre-processed columns corresponding to the triples: "city", "province", "country",
- Both in entire addresses' strings (including the institutions) and in pre-processed columns.

The text file can be compressed before the upload, to speed up the transfer. Once one file is uploaded in NETSCITY, it is possible to import another file without deleting the previously uploaded one. This allows us to deal with a bigger dataset and adding the data progressively, coming from various bibliometric sources. According to users' needs, we are eager to improve the import options and open it to other types of sources and formats.

Geocoding and using the error-correction tool

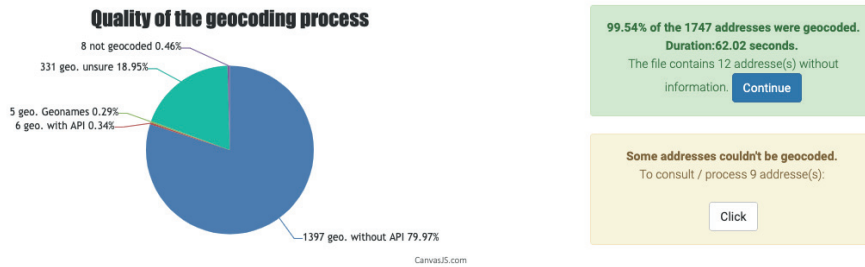


Figure 1. The result of the geocoding process showing a breakdown of the methods used to geocode the addresses from the uploaded addresses

After the import step, the user gets a report detailing the quality of the geocoding process. In Figure 1, the geocoding process has been applied to a homemade CSV file. The bibliometric records included in this file are coming from different sources: Web of Science, Scopus and a French open archive (HAL). As a result, the quality of the geocoding process is poorer than when applied on Web of Science or Scopus files only. However, nearly all the 1,747 addresses included in this file were geocoded. The majority were geocoded with no recourse to an API[3], which means that our internal geocoding knowledge base was sufficient to resolve the

[2] Depending on how you use Scopus, this might be displayed according to your export preference. To select a different export method than what is displayed, click on the drop-down arrow
[3] An external service providing geocoding information.

geographic coordinates. The remaining addresses were geocoded using either a state-of-the-art gazetteer (Geonames), integrated in the application, or an online geocoding API (LocationIQ service).

Identifiant	Adresse	City	Province	Country	Latitude	Longitude	
WOS:000426976100062	VTU, BELAGAVI, INDIA	BELAGAVI		INDIA			Save
WOS:000387985000007	SAVOIE TECHNOLAC, F-73373 LE BOURGET DU LAC, FRANCE	LE-BOURGET-DU-LAC	LAC	FRANCE			
WOS:000366848100332	GRAD SCH ENGN, KAGAWA 7610396, PEOPLES R CHINA			CHINA			
WOS:000374164900016	EXOMARS PROJECT, NL-2201 AZ NOORDWIJK, NETHERLANDS	NOORDWIJK	AZ	NETHERLANDS			
WOS:000380485500366	ENVIRONM OCCUPAT & AGEING PHYSIOL LAB, IXELLE, BELGIUM			BELGIUM			
WOS:000380550000310	DEPT INFORMAT & SISTEMAS, LAS PALMAS DE GC 35017, SPAIN	LAS-PALMAS-DE-GC	GC	SPAIN			
WOS:000367486600022	UJI, CASTELLN 12006, SPAIN			SPAIN			
WOS:000366848100332	FAC ENGN, KAGAWA 7610396, PEOPLES R CHINA			CHINA			

Figure 2. The error-correction interface allowing users to amend the results of the geocoding process

As for non-geocoded addresses, NETSCITY offers the possibility to type some missing spatial information or corrections to complete the geocoding process (Figure 2). Verifying or filling the content of the fields in red (i.e., “city”, “province” and “country”) enhances the geocoding process. If the geocoding fails despite these complements or corrections, it is still possible to add manually the geographical coordinates corresponding to the non-geocoded addresses. In a future version of NETSCITY, we plan to let users modify the result of the geocoding process for geocoded addresses to thwart false positives.

After this first step, users can export the result of the geocoding process, add another file or directly go to the data exploration tools. When choosing to export the results of the geocoding process, NETSCITY generates a list of addresses together with the eventually corrected name of the corresponding cities, provinces and countries (these are sometimes misspelled), the publications ID and the geographic coordinates in a CSV file.

Exploring the dataset through statistical tables and maps

Four types of data exploration are available. The users are presented with:

- Production data through a map and a table listing the number of publications per spatial entity (urban area/country)
- Collaboration data through a map and a table listing the number of collaborations between all pairs of spatial entities (urban areas/countries)

When accessing these different views, the users can change the counting method and unit so that they can immediately see the effects of their choice on the ranking (table view) and on the visualisation (map view).

In the current version, two levels of counting units are available: the users can opt for 1) counting the number of addresses per spatial entity (urban area/country) or 2) counting the number of urban areas or countries. Then, the users are free to choose between full and fractional counting methods. Applied to collaboration data at the urban area level, the fractional counting method implies that by adding the weights of all the links, we obtain the total number of co-written publications of the dataset. It implies weighting the links according to the number of interurban links per co-publication. For instance, if a given publication stems from three different urban areas, each inter-urban link receives 1/3 as a weight for this publication. More generally, if a publication is co-signed from n urban areas, each pair of urban areas (A, B) with A < B is assigned a value l equals to: $1/n(n-1)/2 = 2/(n(n-1))$.
When exploring the numerical data online, all temporal information is overlooked. Longitudinal analysis is available at the export stage, however: the option "with the evolution by years" computes the different variables (number of publications and number of collaborations) on an annual basis. Then, the export final step allows, on the one hand, opting for full or fractional counting and, on the other hand, selecting between CSV or JSON formats.

Availability of NETSCITY as an open source application

The source code is to be released in an online open-source forge (like GitHub) with a suitable license (like one compatible with the well-used GNU-GPL), after several enhancements and another pass of testing and debugging. The project is still ongoing, and we are eager to collect feedback and ideas to expand its capabilities, to tune it to new needs, and to enhance interoperability with other analysis applications such as the CorTexT project[4] or VOSViewer. Together with several bibliometrics tools, the geocoding tool and the aggregation tool included in NETSCITY could be available as off-the-shelf modules.

Geographical underpinning of a set of bibliographic references: a case study using NETSCITY

To test NETSCITY on a specific set of publications, we decided to focus on the publications referring in their titles, abstracts or keywords to the use of a common device: ROV or AUV, which is a submarine robot. Therefore, we issued the following topic search in the *Web of Science Core Collection*: TS= ("ROV" OR "AUV") AND "robot*" AND "underwater", all publication years combined. This query returned 1,004 records published between 1991 and 2017 (Figure 3). We consider this WOS result in the running example that follows.

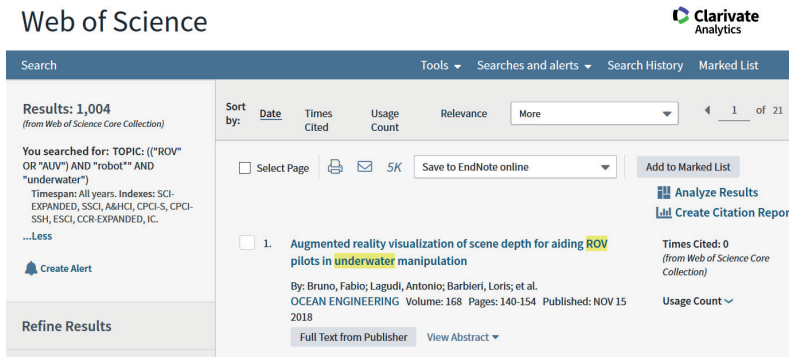


Figure 3. Result page for a sample Web of Science query with 1,004 records

[4] https://www.cortext.net

Although the *Web of Science* interface cannot reveal the leading urban areas producing knowledge on the matter under consideration, NETSCITY lists the top 10 urban areas which publish the most about ROV/AUV (Table 1) at once.

Table 1. Top ten urban areas which publish the most in the ROV/AUV field (Web of Science)

Urban area	*Country*	*Full number of publications*
Tokyo	Japan	44
Boston	United-States	41
Genoa	Italy	41
Shanghai	China	39
Girona	Spain	37
Pisa	Italy	30
Florence	Italy	28
Woods-Hole	United-States	26
Edinburgh	United-Kingdom	24
Singapore	Singapore	23

Then, we can identify the top 10 most intense interurban collaborations about ROV/AUV according to a different counting method (Table 2).

Table 2. Top ten most intense interurban collaboration in the ROV/AUV field (Web of Science)

Urban area 1	*Country 1*	*Urban area 2*	*Country 2*	*Fractional number of collaborations*
Florence	Italy	Pisa	Italy	30.0
Waterloo-Guelph	Canada	Shanghai	China	16.0
Genoa	Italy	Pisa	Italy	14.0
Florence	Italy	Genoa	Italy	9.5
Geelong	Australia	San-Antonio	United-States	9.0
Oslo	Norway	Trondheim	Norway	8.5
Kitakyushu	Japan	Tokyo	Japan	8.0
Brest	France	Shanghai	China	7.5
Lisbon	Portugal	Porto	Portugal	7.0
Woods-Hole	United-States	Baltimore	United-States	6.5

In addition, we can interactively explore the world production map and the world collaboration map per urban area and per country using different counting methods. Figure 4 and Figure 5 are screenshots taken from NETSCITY cartographic zoomable views.

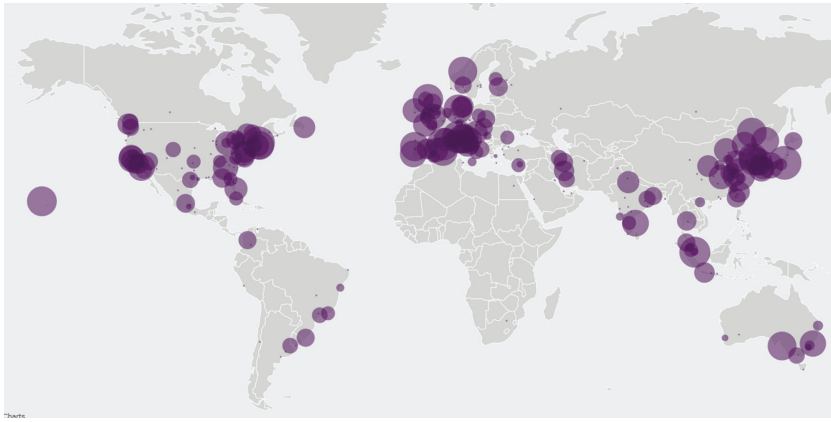


Figure 4. Production map for the 1,004 WOS records of the running example, after geocoding and aggregation are performed. Fractional counts at the urban area level.

From these four views (we could also have chosen to discuss the networks views available in NETSCITY), we can notably observe that the ROV/AUV topic involves publications coming from all over the world, but mainly from maritime cities (Figure 5). We can also notice an important Italian cluster of scientific collaborations on this research issue (Table 2). It is worth noting here that we obtained all these results with a few clicks only. NETSCITY constitutes an effective application to efficiently and quickly visualise and explore the geography of a bibliometric dataset according to different levels of analysis and counting methods.

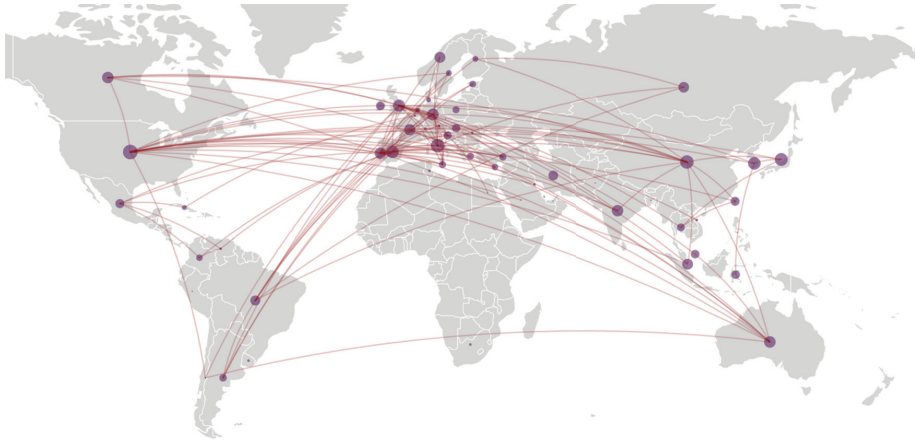


Figure 5. Collaboration map for the 1,004 WOS records of the running example, after geocoding and aggregation are performed. Fractional counts at the country level.

Conclusion

The field of bibliometrics already develops and provides a variety of software to process bibliometric data without requiring any programming skills. However, most existing software fail to account for the geographical complexity of bibliometric data, which leads to misleading results. In addition, many applications generate graph-based maps of locations but, to the best of our knowledge, there is no fully integrated solution to parse and clean the affiliations recorded in a set of references, geolocalising them, clustering them at the requested level of

analysis, and mapping them both on a world base map and in a relational space. The goal of NETSCITY is to offer and promote such a solution.

Concisely, NETSCITY is a new application designed for the general public to perform spatial bibliometrics at the click of a mouse. It computes metrics and produces maps abiding by the state-of-the-art methods of geography of science. We wish that NETSCITY contributes to overcome the pitfalls of spatial data and better inform policy makers regarding territorialised policies worldwide.

Acknowledgments

We would like to thank our entire geography of science research team, and especially Béatrice Milard, Michel Grossetti, and Denis Eckert. This research is supported by LABEX SMS (ANR-11-LABX-0066) under project codenamed Netscience.

References

Andurand, A., Jégou, L., Maisonobe, M., & Sigrist, R. (2015). Les mondes savants et leur visualisation, de l'Antiquité à aujourd'hui. *Histoire et Informatique*, 18/19, 59–94.

Bach, B., Riche, N. H., Fernandez, R., Giannisakis, E., Lee, B., & Fekete, J.-D. (2015). NetworkCube: bringing dynamic network visualizations to domain scientists. In *Posters of the Conference on Information Visualization (InfoVis)*.

Börner, K. Atlas of Science: Visualizing What We Know. Cambridge, MA: MIT Press.

Bornmann, L., & de Moya-Anegón, F. (2018). Spatial bibliometrics on the city level. *Journal of Information Science*, 0165551518806119. https://doi.org/10.1177/0165551518806119

Boyack, K. W., Klavans, R., & Börner, K. (2005). Mapping the backbone of science. *Scientometrics*, 64(3), 351–374. https://doi.org/10.1007/s11192-005-0255-6

Chen, C. (2006). CiteSpace II: Detecting and visualizing emerging trends and transient patterns in scientific literature. *Journal of the American Society for Information Science and Technology*, *57*(3), 359–377. https://doi.org/10.1002/asi.20317

Chen, C. (2016). *CiteSpace: A Practical Guide for Mapping Scientific Literature*. Nova Publishers.

Cobo, M. J., Lopez-Herrera, A. G., Herrera-Viedma, E., & Herrera, F. (2011). Science mapping software tools: Review, Analysis, and Cooperative Study Among Tools. *Journal of the American Society for Information Science and Technology*, 62(7), 1382–1402.

Cobo, M. J., Dehdarirad, T., García-Sánchez, P., & Moral-Munoz, J. A. (2018). Quantifying the reproducibility of scientometric analyses: a case study. In *STI 2018 Conference Proceedings* (pp. 925–933). Leiden: Centre for Science and Technology Studies (CWTS).

Csomós, G. (2018). A spatial scientometric analysis of the publication output of cities worldwide. *Journal of Informetrics*, 12(2), 547–566. https://doi.org/10.1016/j.joi.2018.05.003

Dodge, M., McDerby, M., Turner, M., (2011). Geographic Visualization: concepts, Tools and Applications, Wiley.

Frenken, K., Hardeman, S., & Hoekman, J. (2009). Spatial scientometrics: Towards a cumulative research program. *Science of Science: Conceptualizations and Models of Science*, *3*(3), 222–232. https://doi.org/10.1016/j.joi.2009.03.005

Frenken, K., & Hoekman, J. (2014). Spatial Scientometrics and Scholarly Impact: A Review of Recent Studies, Tools, and Methods. In Y. Ding, R. Rousseau, & D. Wolfram (Eds.), *Measuring Scholarly Impact* (pp. 127–146). Cham: Springer International Publishing. Retrieved from http://dx.doi.org/10.1007/978-3-319-10377-8_6

Gauffriau, M., Larsen, P., Maye, I., Roulin-Perriard, A., & Ins, M. (2008). Comparisons of results of publication counting using different methods. *Scientometrics*, *77*(1), 147–176. https://doi.org/10.1007/s11192-007-1934-2

Hoekman, J., Frenken, K., & Tijssen, R. J. W. (2010). Research collaboration at a distance: Changing spatial patterns of scientific collaboration within Europe. *Research Policy*, *39*(5), 662–673. https://doi.org/10.1016/j.respol.2010.01.012

Jégou, L. (2014). Toward spatially referenced academic data at global scale: the full geocoding of Wos-Datasets, methods and results. Presented at the 2nd Geography of Innovation International Conference, Utrecht.

Leydesdorff, L., & Persson, O. (2010). Mapping the Geography of Science: Distribution Patterns and Networks of Relations among Cities and Institutes. *Journal of the American Society for Information Science and Technology*, *61*(8), 1622–1634. https://doi.org/10.1002/asi.21347

Leydesdorff, L., & Park, H. W. (2017). Full and Fractional Counting in Bibliometric Networks. *Journal of Informetrics*, *11*(1), 117–120.

Maisonobe, M., Jégou, L., & Eckert, D. (2018). Delineating urban agglomerations across the world: a dataset for studying the spatial distribution of academic research at city level. *Cybergeo: European Journal of Geography*, e871. https://doi.org/10.4000/cybergeo.29637

Maisonobe, M. & Jégou, L., (2018). Explorer les réseaux mondiaux : proposition d'outil interactif combinant graphe (diagramme nœuds-liens) et carte de flux. Presented at the CIST2018 - Représenter les territoires / Representing territories, Rouen.

Matthiessen, C. W., Schwarz, A. W., & Find, S. (2010). World Cities of Scientific Knowledge: Systems, Networks and Potential Dynamics. An Analysis Based on Bibliometric Indicators. *Urban Studies*, *47*(9), 1879–1897. https://doi.org/10.1177/0042098010372683

Nomaler, Ö., Frenken, K., & Heimeriks, G. (2014). On Scaling of Scientific Knowledge Production in U.S. Metropolitan Areas. *PLOS ONE*, *9*(10), e110805. https://doi.org/10.1371/journal.pone.0110805

Pan, R. K., Kaski, K., & Fortunato, S. (2012). World citation and collaboration networks: uncovering the role of geography in science. *Scientific Reports*, *2*. https://doi.org/10.1038/srep00902

van Eck, N. J., & Waltman, L. (2009). How to Normalize Co-occurrence Data? An Analysis of Some Well-Known Similarity Measures. *Journal of the American Society for Information Science and Technology*, *60*(8), 1635–1651. https://doi.org/10.1002/asi.21075

Waltman, L., van Eck, N. J., & Noyons, E. C. M. (2010). A unified approach to mapping and clustering of bibliometric networks. *Journal of Infometrics*, *4*(4), 629–635. https://doi.org/10.1016/j.joi.2010.07.002

van Eck, N. J., & Waltman, L. (2010). Software survey: VOSviewer, a computer program for bibliometric mapping. *Scientometrics*, *84*(2), 523–538. https://doi.org/10.1007/s11192-009-0146-3

Van Hooydonk, G. (1997). Fractional counting of multiauthored publications: Consequences for the impact of authors. *Journal of the American Society for Information Science*, *48*(10), 944–945. http://doi.org/d9f6jz

Waltman, L., Tijssen, R. J. W., & Eck, N. J. van. (2011). Globalisation of science in kilometres. *Journal of Informetrics*, *5*(4), 574–582. https://doi.org/10.1016/j.joi.2011.05.003

Mapping Disciplinary Knowledge Flows Using Book Reviews

Alesia Zuccala[1], Helena H. Zhang[2,3] and Fred Y. Ye[2,3]

[1] *a.zuccala@hum.ku.dk*
Dept. of Information Studies, University of Copenhagen, Njalsgade 76, 2300 København S, (Denmark)

[2,3] *137356390@qq.com; yye@nju.edu.cn*
[2]*International Joint Informatics Laboratory (IJIL), Nanjing University – University of Illinois, Nanjing - Champaign, (China - USA)*

[3]*Jiangsu Key Laboratory of Data Engineering and Knowledge Service, School of Information Management, Nanjing University, Nanjing 210023, (China)*

Abstract

In this paper we re-examine the work of Lindholm-Romantschuk (1998), who was the first to study disciplinary knowledge flows using book reviews from the social sciences and humanities. Our study is based on a current, and larger sample of book reviews and includes reviews written for scientific books as well. We use a different approach to classifying book titles, and employ different indicators for measuring knowledge INflows and OUTflows. For most book reviewing disciplines, a book is likely to receive on average one or two reviews across a five-year period. Some disciplines have books that are reviewed ≥10 times, and it is with this 'elite' sample that we develop VOSviewer knowledge flow maps, and measure Flow Ratios (FR) as well as OUT/IN Ratios (OI). Our maps show where the strongest overall knowledge flows exists amongst different disciplines. We also show the degree to which book reviewing disciplines possess *permeability of boundaries*. Those with a lower degree of permeability (e.g., *History*) tend to be highly independent. They may take in reviews, but do not 'need' other disciplines for external reviews, while those with a balanced INflow-OUTflow tend to exchange reviews a lot with other disciplines (e.g., *Linguistics; Sociology; Communication studies*).

Introduction

The most informative work to date concerning scholarly book reviews is Ylava Lindholm-Romantschuk's (1998) monograph: "*Scholarly Book Reviewing in the Social Sciences and Humanities. The Flow of Ideas Within and Among Disciplines."* Originally it was written as a PhD thesis, and conceived later as a monograph to show "that scholarly book reviews are significant indicators of scholarly communication, and can successfully be utilized to trace the flow of information within and across knowledge domains" (p, viii).

In the latter part of the 1990s, this research began with as "an elite sample" of monographs published by university presses between 1971 and 1990. All of the book titles were taken from the *Outstanding Academic Books* list, now known as the *CHOICE Outstanding Academic Titles*. Using the Bibliolinks software, scholarly reviews were matched to the *Outstanding* book titles and retrieved from the *Arts & Humanities Search* and *SocialSciSearch* available through the database vender known as *Dialog*. A final dataset of 13,924 reviews was established, and each was then used to examine, and map the inflow of information into a discipline and outflow of information into a discipline. As Lindholm - Romantschuk (1998) explains: "The higher the proportion of reviews appearing in a discipline's journals that are of books originating in other disciplines, the higher the inflow of information into that discipline, and vice versa" (p. 93).

Here, we will re-examine Lindholm-Romantschuk's (1998) work. We use the term 're-examine' because the aim is not a direct reproduction, but to implement an alternative method and propose a new hypothesis concerning knowledge-flows from book reviews. We also assume that what this author found in 1998 will not have changed significantly. In other

words, reviews still diffuse knowledge and ideas to other disciplines, but we expect this transmission to be amplified as a result of newer journals established since that earlier period (1971-1990), including research trends related to more and more inter-disciplinarity.

A newer and larger dataset was therefore retrieved from Clarivate's Web of Science, and similar to this early study, we utilize book reviews published in journals from the *Arts & Humanities Citation Index* and the *Social Sciences Citation Index*, but include the *Science Citation Index* as well. Between the years of 2001 and 2015, a total of 856,474 book reviews were written and published in these three broad categories/fields. With this set of reviews, our study is focused on the following questions:

1. What are the top book disciplines that receive ≥10 reviews per book, and what is the mean or average number of reviews based on the total reviews from the discipline?
2. For every book that has been reviewed ≥10, what is the subject-area flow (note: Similar to Lindholm-Romantschuk (1998) we will examine information ***INflow*** and ***OUTflow*** (note: definitions are provided in our methodological section below)

Related Literature

Book reviews as evaluation tools

Books as well as book reviews have not been featured so often in the literature on research evaluation (Zuccala & Robinson-Garcia, 2018). Either they have been excluded from evaluations or set aside for special consideration, thus reviews have similarly received little attention since they are not original pieces of research. As such, they are secondary to the scholarly communication system, and have a history of being useful to collection development librarians (Blake, 1989; Furnham, 1986; Natowitz & Wheeler Carlo, 1997; Parker, 1989; Serebnick, 1992) or scholars needing quick synopses to make decisions about their academic reading and/or teaching (Spink, Robbins, & Schamber, 1998; Hartley, 2006).

Book reviews are barely ever cited (Diodatao, 1984) and this seems to be true mostly of the Type I genre, compared to Type II, which review several books related to the same topic at the same time (Zuccala & van Leeuwen, 2011). Nicolaisen (2002) has shown; however, that books receiving positive or favorable reviews tend to be cited more often than those receiving neutral or negative comments. Since review counts and book citations tend to correlate positively, Gorraiz et al. (2014) suggest using book reviews to select books for commercial indexes like the Book Citation Index™ or Scopus. According to Zuccala et al. (2014) book reviews may be likened to 'citations', and thus characterized as 'mega-citations'. In this study, sentiment analysis and machine learning techniques were used to classify book on the basis of how they capture a book's scholarly credibility and written quality. By classifying reviews, it is possible then to code them and weight them, so that each book or a set of books might receive an overall 'quality' indicator.

Mapping disciplinary knowledge flows

Book reviews play a facilitating role in the scholarly communication system, as secondary 'knowledge sharing' pieces, and it is in this sense, that they are related to the literature pertaining to disciplinary knowledge flows. Typically knowledge flows are traced using 'citations' and the degree to which one discipline cites research from another discipline (Porter & Chubin, 1985; van Leeuwen & Tijssen, 2000). According to Leeuwen & Tijssen (2000) "scientific knowledge is 'systemic'" and "occurs as an interactive and collective process among interdependent agents and institutions sharing and exchanging information and knowledge" (p. 186). In this sense, the flow of knowledge from research to technological innovation can also be observed utilizing citations to patents (Alcácer & Gittelman, 2006;

Nomaler & Verspagen, 2009). With citations, knowledge flows can be measured at the micro-level between papers, or at the journal level, where journals assigned to different disciplines exchange citations with each other. Here we can see where specific fields recognize the work of other fields, even though the recognition of a journal's core discipline or field is not always so precise. Commercial databases (i.e., WoS and Scopus) assign certain journals to more than one disciplinary category, thus assignment processes can vary (Glänzel & Schubert, 2003).

With book reviews, knowledge flows may be examined at the level of journals as well (Lindholm-Romantschuk, 1998). However, unlike studies using only articles, at least two classifications need to be taken into consideration: 1) the classification of the journal, and 2) the classification of the book, according to its Library of Congress Classification or Dewey Decimal Code. For example, the book *Athena unbound: The advancement of women in science and technology* classified as LCC Q130.E85 (i.e., women in science; women in technology) has a review published in the journal *Technological Forecasting and Social Change* (WoS Categories: Business; Planning & Development). Here we see a flow from the book discipline of women in science and technology studies to business, in addition to planning and development.

Originally Lindholm-Romantschuk (1998) mapped this type of knowledge flow using a very small sample of titles (i.e., 13,924 reviews of 1,657 books). Here we start with a significantly larger selection of book reviews (*N*=856,474 reviews of *N*=510,692 books), and work with titles that do not have a Library of Congress Classification (LCC) code, since the Web of Science does not include this information. A separate, onerous procedure might be employed to collect the individual LCC codes per book (perhaps via WorldCat); however, we have chosen instead to implement a different method for classifying each book, outlined in the next section.

Methodology

Data Collection

Our dataset of book reviews was collected from the Web of Science, based on the following search strategy: Language=English, Type=Book Review, Databases=Science Citation Index (SCI-E) + Social Sciences Citation Index (SSCI) + Arts & Humanities Citation Index (AHCI). Again, for the publication years (PY) of 2001 to 2015 we retrieved a total of 856,474 book reviews from the Web of Science. In order to test for and observe changing knowledge flow patterns across these 15 years we collected separate datasets for three comparative 5-year time periods: 1) 2001 to 2005, 2) 2006 to 2010, and 3) 2011 to 2015. Table 1 presents how the initial dataset was divided according to the three time periods. To process the data further for this study, we also chose to focus on books that had received equal to or more than ten reviews (≥10).

Table 1. Initial dataset (*N*=856,474) divided according to the three time periods.

Time period (PYs)	Number of reviews	Number of books	Number of books receiving ≥ 10 Reviews
2001-2005	290149	175576	689
2006-2010	286917	169990	776
2011-2015	279408	165126	686
	N=856,474	***N=510,692***	

Tables 2, 3, and 4 below, presents the top ten Web of Science categories (WC) containing ≥10 book reviews per book (2001-2005; 2006-2010; 2011-2015), versus the average number of reviews per book taken from the TOTAL number of reviews for this category.

Table 2. Top 10 categories with ≥10 reviews and average reviews per book (2001-2005)

Categories	Total # Books	Total # Book Reviews	Average reviews per book	Books with ≥10 reviews
History	30630	54866	2	414
Humanities, Multidisciplinary	21706	30431	1	48
History & Philosophy Of Science	4007	6871	2	41
Political Science	7260	12042	2	37
Philosophy	4104	6977	2	27
Literature	9722	13484	1	21
Sociology	5552	9483	2	19
Economics	4606	8088	2	17
Religion	9575	15566	2	14
Psychiatry	3464	5494	2	13

Table 3. Top 10 categories with ≥10 reviews and average reviews per book (2006-2010)

Categories	Total # Books	Total # Book Reviews	Average reviews per book	Books with ≥10 reviews
History	29458	54915	2	425
Humanities, Multidisciplinary	21860	32032	1	73
History & Philosophy Of Science	4214	7388	2	55
Political Science	7741	13806	2	50
Sociology	5709	9968	2	36
Anthropology	4066	6881	2	26
Religion	11634	18949	2	23
Literature	7549	11060	1	23
Economics	3863	6698	2	20
Geography	2380	4193	2	20

Mapping knowledge flows

To map the knowledge flows between books and journals we prepared the following classification scheme. For each journal, the ***Web of Science category (WC)*** was used, but for books, we developed a classification based on a method of statistical induction. Again, this needed to be done in the absence of the book's LCC Classification. First, we listed all journals that published reviews on a specific book, then to give the book itself a classification, we chose the corresponding Web of Science code, which represented the majority of the published reviews (***WC => BD***).

In cases where it was possible to detect a 'majority' of ***Web of Science Categories (WC),*** we gave the book multiple classifications at the same time (i.e., multiple **BDs**). Table 5, shows an example of how this can occur, with the book "***A New History of the Humanities: The Search for Principles and Patterns from Antiquity to the Present***". Note

that this title has been reviewed in five different journals, and appears to be linked to a variety of ***WC*** codes, thus leading to more than one classification (***BDs***).

Table 4. Top 10 categories with ≥10 reviews and average reviews per book (2011-2015)

Categories	**Total # Books**	**Total # Book Reviews**	**Average reviews per book**	**Books with ≥10 reviews**
History	28708	54895	2	408
History & Philosophy Of Science	3837	6849	2	52
Political Science	8428	14998	2	48
Humanities, Multidisciplinary	21453	30573	1	32
Sociology	5975	10602	2	31
Anthropology	3967	6966	2	31
Religion	13846	22906	2	26
Literature	7550	11073	1	24
Area Studies	7117	12110	2	14
International Relations	4115	7352	2	10

Table 5. Example of a 'multidisciplinary' book assigned to many WCs

Book Title	Journal with a Review	Web of Science Category (WC)
Bod, R. (2013). A New History of the Humanities: The Search for Principles and Patterns from Antiquity to the Present	*ISIS*	History & Philosophy of Science
	Renaissance Quarterly	Medieval & Renaissance Studies
	American Historical Review	History
	Postmedieval-A Journal of Medieval Cultural Studies	Cultural Studies
	TLS - Times Literary Supplement	Humanities, Multidisciplinary

Next, in order to ensure a 'robust' mapping we focus on books that have been reviewed equal to or more than (≥ 10) and utilize the following definitions of ***INflow*** and ***OUTflow:***

a. ***INflow*** = the amount of book reviews published in a *Web of Science Category (WC)* journal originating from other disciplines *(BD). (e.g., a review of an economics book in a history journal is one inflow from economics to the discipline of history)* *(BD =>* ***WC <= BD****)*

b. ***OUTflow*** = the amount of book reviews related to a specific book discipline *(BD)* that are published in *Web of Science Category* (WC) journals from outside that discipline (e.g., *A review of a history book in another disciplinary journal is counted as one unit of outflow from the discipline of history*). *(****BD => WC*** *<= BD)*

Based on all reviews collected for the full period of 2001 to 2015, we identified a total number of 126 Web of Science Categories (*WCs*) and found 62 unique book disciplines (*BDs*). The number of WC and BD differed for the separate five year time periods, where in 2001 to 2005 the BD=53 and the WC=112; in 2006 to 2010 the BD=51, and the WC=107; In 2011 to 2015, the BD=45 and the WC=105.

[*NOTE: We didn't process the data for the full period of 2001-2015. We obtain the total number through the combination of WC and BD in 2001-2005, 2006-2010 and 2011-2015 and then remove the duplicate data.]*

We can therefore calculate the ***INflow*** to the Web of Science Category journals (*WC*) from book disciplines (*BD*) and the ***OUTflow*** from the book disciplines (*BD*) to Web of Science category journals (*WC*).

An outflow – inflow ratio, defined overall as the Flow Ratio (FR) is measured as follows:

$$FR = \frac{OUT - IN}{IN} \qquad (1)$$

The second measure, which we refer to as the OUT/IN ratio (OI) indicates the relative values of OUT versus IN, and may be defined as:

$$OI = \frac{OUT}{IN} \qquad (2)$$

Both FR and OI mark the "knowledge flows" of book reviews, whereby the larger the FR and OI, the higher the review flows. An FR > 0 or an OI > 1 indicates that the knowledge output via book review is larger than its input. The relatively higher FR or OI reveals the higher outflow disciplines, which means that the books from these disciplines are more often reviewed in Web of Science categories outside their own discipline.

For the disciplines with FR < 0 or OI < 1, it is true that these disciplines "take in" more inflow from other disciplines. Note that we always discuss knowledge "flow" (outflow or inflow) among disciplines via book reviews in this paper. Hence, an FR > 0 and an OI > 1 indicates that the knowledge outflow in any discipline is larger than its inflow.

After calculating the average number of book reviews in the discipline of History and selecting all books that had been reviewed equal to or greater than ten (≥10) times we show via the VOSViewer (version 1.6.9) tool how one *INflow* and one *OUTflow* illustrates the portion of reviews exchanged between a Book Discipline *(BD)* and a Web of Science journal category (*WC*). As part of our data processing, Python coding was used in addition to VOSViewer.

Results

Figures 1, 2 and 3, for the time periods of 2001 to 2005, 2006 to 2010, and 2011 to 2015 respectively present VOSviewer maps of the overall disciplinary knowledge *INflows* and *OUTflows.* Again, the term *INflow* refers to the flows going into the Web of Science Category journals (*WCs*) from the book disciplines (*BD*) and *OUTflow* means flows from the book disciplines (*BD*) to the Web of Science category journals (*WC*).

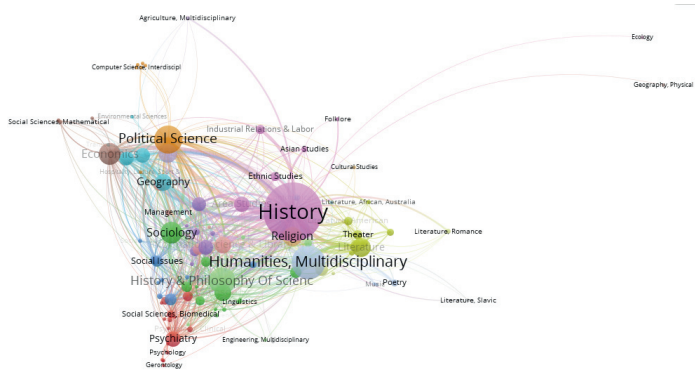


Figure 1. Overall flows in 2001-2005
(53 book disciplines and 112 WCs of journals with reviews)

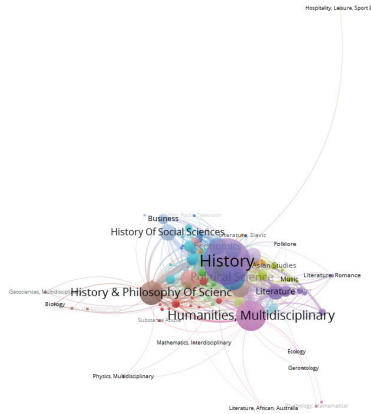


Figure 2. Overall flows during 2006-2010
(51 book disciplines and 112 WC of journals with reviews)

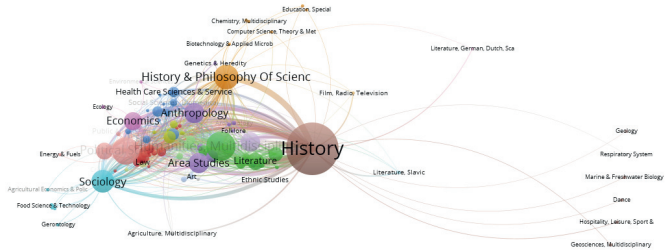


Figure 3. Overall flows during 2011-2015
(45 book disciplines and 105 WC journals with reviews)

Below, we present three VOSviewer maps, where Figure 4 (2011-2015), Figure 5 (2011-2010) and Figure 6 (2011-2015) illustrate the total *INflows* and *OUTflows* for the disciplines of *Archaeology*, *History*, and the *Health Sciences*. In all Figures, the lines represent the absolute sum of *INflow* and *OUTflow*, with the shaded (less visible) lines indicating where there is no flow with other existing disciplines at the level of reviews ≥10 and weights ≥10.

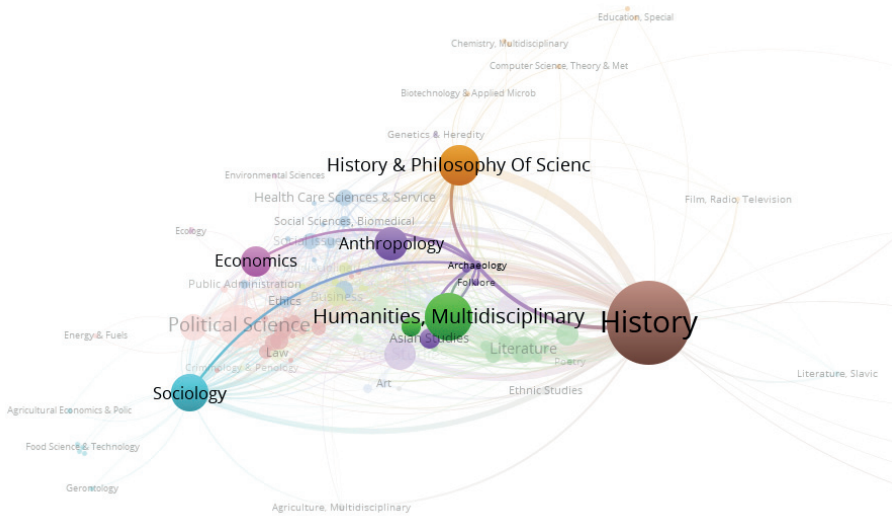


Figure 4. IN and OUT knowledge flows for Archaeology (2011-2015)

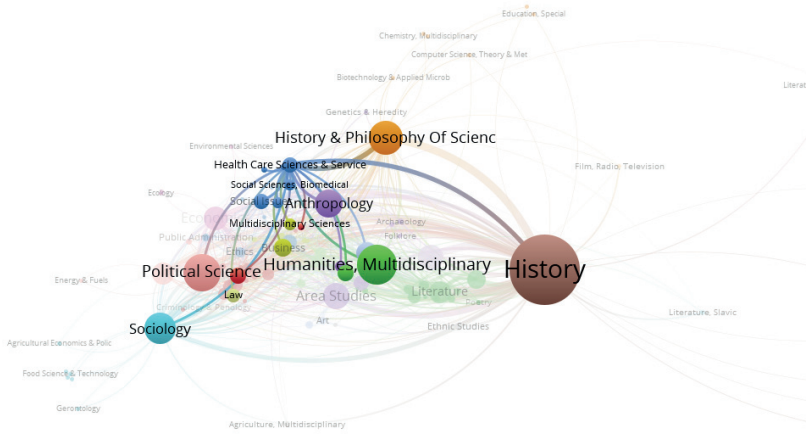


Figure 5. IN and OUT knowledge flows of Health Sciences (2011-2015)

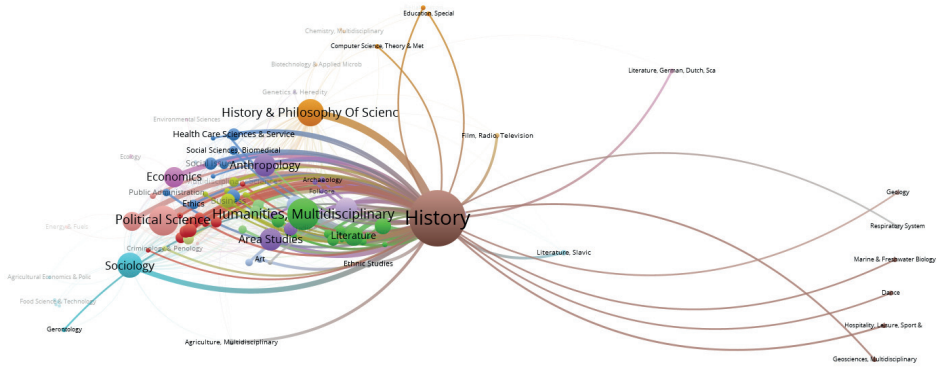


Figure 6. IN and OUT knowledge flows for History (2011-2015)

When we apply both the FR and OI as overall measures, the relative distribution of ***INflow*** and ***OUTflow*** may be calculated for all disciplines. Thus, Figures 7, 8 and 9 below show the global distribution of all disciplines, where an FR > 0 and OI >1 indicates that the ***OUTflow > INflow***.

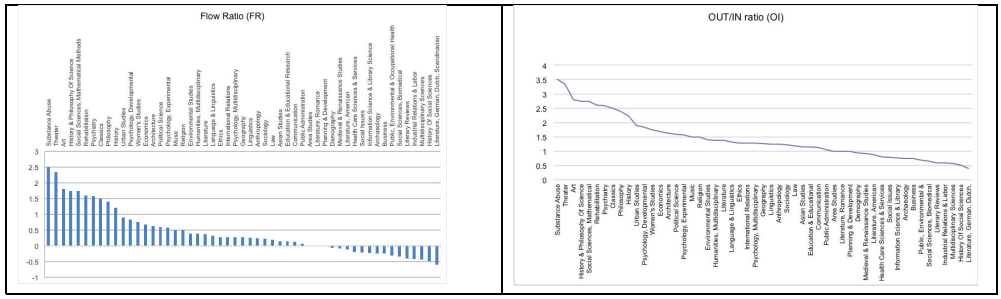


Figure 7. The Flow Ratio (FR) plot and OUT/IN (OI) curve for the period of 2001-2005

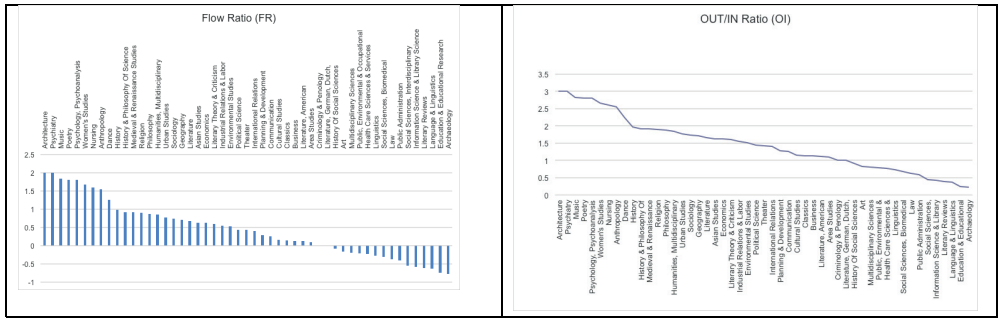


Figure 8. The Flow Ratio (FR) plot and OUT/IN (OI) curve for the period 2006-2010

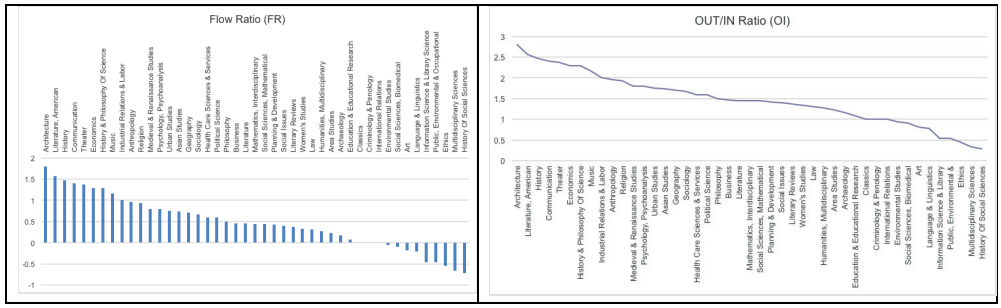


Figure 9. The Flow Ratio (FR) plot and OUT/IN (OI) curve for the period 2011-2015

Conclusions

A few conclusions may be drawn from this research. Earlier, Lindholm-Romantschuk (1998) found that "the mean number of reviews per book varies across disciplines, from 3.75 (communication) to 10.57 (history and geography) with the mean for all books being 7.77 reviews" (p. 53). With Tables 2, 3, and 4 we show that the average number of book reviews recorded in journals indexed in the Web of Science is approximately 1 or 2 reviews per book. These average values differ significantly from what Lindholm-Romantschuk found, but this is may be related to the fact that she chose to work with a much smaller elite sample of Choice reviews. In most Web of Science, journals few books will be reviewed greater than or equal to ten (≥10) times; however for this study we have chosen it to be the cut-off point for our own 'elite' sample.

Across the different time periods, there are similarities with respect to the type of disciplines that produce the most frequent book reviews, where *History*, *Humanities Multidisciplinary*, *History and Philosophy of Science*, and *Political Science* are consistently listed at the top in all three tables across the three five-year time periods (i.e., Tables 2, 3, and 4). Although books from these disciplines are indeed published and reviewed quite frequently, to some extent the numbers featured in our tables are an artifact of the WoS database, as they depend on the type of disciplinary journals that have been indexed.

The static maps shown in this paper are not ideal for interpretation, but when the dynamic features of VOSViewer are used to examine, for instance, the overall flows for all disciplines (2011-2015) *History* is observed to be the most dominant field with strong knowledge flow exchanges with a variety of other disciplines, like *Literature, Political Science, Economics, Anthropology, Sociology, Communications, Area Studies, Business*, etc. This is to be expected, and can be contrasted with another discipline, like *Archaeology*, which exchanges knowledge via a shared approach to book reviewing only with *History* and *Anthropology*.

Since the mapping approach can only present the total sum of IN and OUT flows, we also employ a Flow Ratio (FR) and OUT/IN ratio (OI) to indicate differences between the disciplines, where some are more open in terms of ***INFlow*** and ***OUTflow*** and others tend to be more closed. Lindholm-Romantschuk (1998) characterizes this in terms of "*permeability of boundaries*" (p. 96).

What Figures 7, 8 and 9 show is that the disciplines at the top of the Flow Ratio (FR) and OUT/IN (OI) scales (e.g., *Art, Theatre, Substance Abuse, Social Sciences-Mathematical Models*) publish books that are often reviewed in journals outside their own discipline (i.e., OUTflow > INflow). These disciplines do not have as many of their own reviewing journals (or at least journals indexed in the WoS); hence, we see more prominent external knowledge

flows. The disciplines at the bottom of the scale posses both low OUTflows and INFlows, and they are also of very low permeability, in that they rarely ever take-in or receive outside reviews from other disciplines.

It is the disciplines that fall within the upper-middle range where the OUT/IN Ratio (OI) ranges from 1.50 to 2.50, which are the most interesting. Similar to Lindholm-Romantschuk (1998), we see that disciplines like *Architecture, Economics, History* are moderately permeable in the sense that outside disciplines often benefit or see the potential utility of the information that they generate via book reviews; yet they do not necessarily "need" these other disciplines to review their books having their own review journals for this purpose. C*ommunication, Literature,* and *Education & Educational Research*, *Ethics, Linguistics, Sociology,* and *Psychology-Multidisciplinary*, are disciplines with higher levels of permeability. Because other disciplines write reviews for their books and they also take-in knowledge from other disciplines they possess fairly balanced knowledge *INflows* and *OUTflows.*

Overall, our findings demonstrate that the extraction of new data pertaining to book reviews and new measures can be and have been useful for re-examining Lindholm-Romantuschuk's (1998) earlier research. However, similar to her approach in the 1990s we have focused on an elite sample of titles receiving ≥10 reviews. Since the average review rates for books is close to one or two reviews within a five-year period, it could be more useful to study book reviewing across disciplinary boundaries based on a more 'average' or normal perspective. We will further this research with a study sample of books receiving greater than or equal (≥5) reviews and less than ten (<10), to see if the overall disciplinary knowledge flows and permeability of boundaries remain similar.

References

Alcácer, J. & Gittelman, M. (2006). Patent citations as a measure of knowledge flows: The influence of examiner citations. *The Review of Economics and Statistics, 88(4),* 774-779.

Barber, B. (1961). Resistance by scientists to scientific discovery. *Science, 134*, 596-602.

Blake, V. (1989). The role of reviews and reviewing media in the selection process: An examination of the research record. *Collection Management, 11(1/2),* 1–40.

Börner K, Klavans R, Patek M, Zoss AM, Biberstine JR, Light RP, et al. (2012) Design and Update of a Classification System: The UCSD Map of Science. *PLoS ONE* 7(7), e39464. https://doi.org/10.1371/journal.pone.0039464.

Diodato, V. (1984). Impact and scholarliness in arts and humanities book reviews: A citation analysis. *Proceedings of the 47th annual meeting of the American Society for Information Science, 21*, 217–221.

Furnham, A. (1986). Book reviews as a selection tool for librarians: Comments from a psychologist. *Collection Management, 8(1),* 33–43.

Glänzel, W. & Schubert, A. (2003). A new classification scheme of science fields and subfields designed for scientometric evaluation purposes. *Scientometrics, 56(3),* 357-367.

Gorraiz, J., Gumpenberger, C. & Purnell, P.J. (2014). The power of book reviews: a simple and transparent enhancement approach for book citation indexes. *Scientometrics, 98(2),* 841-852.

Hartley, J. (2006). Reading and writing book reviews across disciplines. *Journal of the American Society for Information Science and Technology, 57(9),* 1194–1207

Leydesdorff, L., & Rafols, I. (2011). Indicators of the interdisciplinarity of journals: Diversity, centrality, and citations. *Journal of Informetrics*, 5(1), 87–100.

Leydesdorff, L., & Rafols, I. (2012). Interactive overlays: A new method for generating global journal maps from Web-of-Science data. *Journal of Informetrics*, 6(3), 318–332.

Leydesdorff, L., Rafols, I. & Chen, C. (2013). Interactive Overlays of Journals and the Measurement of Interdisciplinarity on the Basis of Aggregated Journal–Journal Citations. *Journal of the American Society for Information Science and Technology*, 64(12), 2573–2586.

McIntosh R P. (1967).An index of diversity and the relation of certain concepts to diversity. *Ecology*, 48:392–404.

Natowitz, A., & Wheeler Carlo, P. (1997). Evaluating review content for book selection: An analysis of American history reviews in Choice, American Historical Review, and Journal of American History. *College & Research Libraries, 58(4)*, 322–335.

Nicolaisen, J. (2002). The scholarliness of published peer reviews: a bibliometric study of book reviews in selected social science fields. *Research Evaluation, 11(3),* 129-140.

Nomaler, ö., & Bart Verspagen, B. (2009). Knowledge Flows, Patent Citations and the Impact of Science on Technology. *Economic Systems Research, 40(4),* 339-366.

Parker, J.M. (1989). Scholarly book reviews in literature journals as collection development sources for librarians. *Collection Management, 11(1&2),* 41–57.

Porter, A. L., & Chubin, D. E. (1985). An indicator of cross-disciplinary research. *Scientometrics, 8*(3-4), 161-176.

Porter, A. L., Cohen, A. S., Roessner, J. D., & Perreault, M. (2007). Measuring researcher interdisciplinarity. *Scientometrics, 72*(1), 117–147.

Rafols, I. & Meyer, M. (2010). Diversity and network coherence as indicators of interdisciplinarity: case studies in bionanoscience. *Scientometrics. 82*(2), 263-387.

Serebnick, J. (1992). Selection and holdings of small publishers' books in OCLC libraries: A study of the influence of reviews, publishers and vendors. *Library Quarterly, 62(3)*, 259–294.

Small, H. (2003). Paradigms, Citations, and Maps of Science: A Personal History. JASIST, 54(5): 394–399.

Small, H.G. & Griffith, B.C. (1974). The structure of scientific literatures: Identifying and graphing specialties. *Science Studies, 4*, 17–40.

Spink, A., Robbins, D., & Schamber, L. (1998). Use of scholarly book reviews: implications for electronic publishing and scholarly communication. *Journal of the American Society for Information Science, 49(4)*, 364–374.

Steele, T. W., & Stier, J. C. (2000). The impact of interdisciplinary research in the environmental sciences: A forestry case study. *Journal of the American Society for Information Science, 51*(5), 476-484.

Stirling A.(2007). A general framework for analysing diversity in science, technology and society. *Journal of the Royal Society Interface*, 4(15): 707-719.

Tijssen, R. J. W. (1992). A quantitative assessment of interdisciplinary structures in science and technology: Co-classification analysis of energy research. *Research Policy, 21*(1), 27–34.

Van Leeuwen, T. & Tijssen, R. (2000). Interdisciplinary dynamics of modern science: analysis of cross-disciplinary citation flows. *Research Evaluation*, 9(3), 183–187.

Zuccala, A. & Garcia, N.R. (2018). Reviewing, indicating, and counting books for modern research evaluation system. in Glanzel, W., Moed, H.F., Schmoch U., Thelwall, M. (ed.). Springer Handbook of Science and Technology Indicators, Springer (Forthcoming).

Zuccala, A. & van Leeuwen, T. (2011). Book reviews in humanities research evaluations. *Journal of the Association for Information Science and Technology, 62(10),* 1979-1991.

Zuccala, A., van Someren, M., & van Bellen, M. (2014b). A machine-learning approach to coding book reviews as quality indicators: towards a theory of megacitation. *Journal of the Association for Information Science and Technology, 65(11)*, 2248-2260

Collaboration Size and Citation Impact in Big Data Research

Xiaozan Lyu[1,2], Xiaojing Cai[1] and Ping Zhou[1]

[1] *lvxz1991@zju.edu.cn; cai_xj@zju,edu.cn; pingzhou@zju.edu.cn*
[1] Zhejiang University, School of Public Affairs, Dept of Information Resources Management, Hangzhou (China)

[2] Leiden University, Center for Science and Technology Studies (CWTS), Kolffpad 1, P.O. Box 905, Leiden 2300 AX (The Netherlands)

Abstract

With rapid and widespread application, Big Data has attracted growing attention from academic communities. Concerning the importance of collaboration in scientific research, the current study focused on profiles of collaboration types and size, as well as their relations with citation impact based on publications indexed in the Web of Science (WoS) from 2003 to 2017, so as to provide decision-making basis for optimizing collaboration options. Results show that collaborated publications, as well as collaboration size (*i.e.*, number of authors, institutions and countries) in Big Data have grown substantially. Within-institutional collaboration plays a main role, while international collaboration takes relatively low but rising share. Citation impact is dependent on collaboration types with international collaboration the most efficient, followed by domestic one. None-collaboration receives the least citations. Team size increases over time in general, and is positively related with citation impact, and such phenomenon exists in different collaboration types. Among all the research fields involved, Mathematics and Computer Science is the main contributor in publications and also benefits the most in collaboration despite of its limited team size.

Introduction

Emerged in the 1990s and gained momentum in the early 2000s, Big Data has been affecting every aspect of our life (Kitchin, 2014) . According to the statistics of popularity from Google Search, Big Data has attracted growing interests especially after 2011.

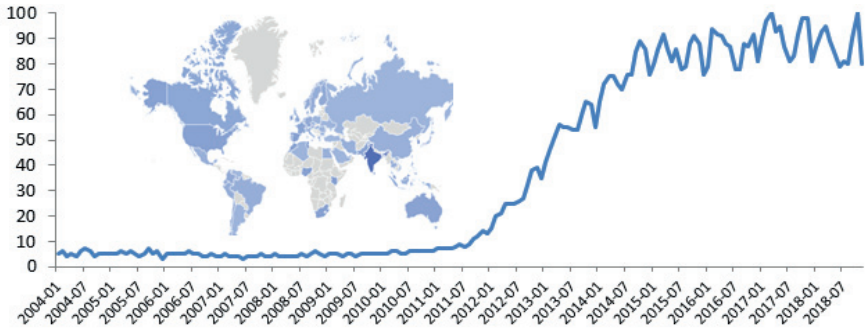


Figure 1. Popularity of term "Big Data" on Google search (Source: Google trends. Retrieve date: 2018-11-06.)

Regarded as "the new oil", the annual global revenue for Big Data market in 2017 has reached $33.5 billion, and the number is expected to be doubled in the next four years[1]. More leading technology companies, such as IBM, SAP, Oracle, Hewlett Packard, and Accenture, are taking the initiative to join the international competition, not only to enhance competitiveness through merger and acquisition integration, but also launch data analysis and other related products and services by accelerating research and development innovation.

It is evident that Big Data has also drawn huge attention from many nations in recent years because of its potential value for increasing business productivity and breakthroughs in scientific research (Chen & Zhang, 2014). The United States first announced the *Big Data Research and Development Initiative* in 2012, committing more than $200 million to Big Data

research projects. Many other countries or regions jointed the campaign in the coming years. For example, the *Digital Agenda for Europe and Challenges for 2012*, the UK data capability strategy: *seizing the data opportunity*, the *Australian Public Service Big Data Strategy*, and the *Planning for the Development of Big Data Industry (2016-2020)* of China.

At the same time, due to the rapid globalization of science and improved technologies, research has progressed through three ages - the individual, the institutional and the national, and is entering a fourth age driven by international collaborations between elite research groups (Adams, 2013). There is an increasing trend for countries to collaborate with each other of dissimilar scientific, economic, and social stature (Hsiehchen, Espinoza, & Hsieh, 2018; Clarke & Plume, 2011; Wagner *et al.*, 2017), as a means of leveraging domestic research capabilities and taking advantage of R&D investments abroad (European Commission, 2003; Wagner, 2008). The globalization of science also can be reflected in the growing proportion of collaborative papers by multiple scientists (Adams, 2013; National Science Board, 2016).

The importance of scientific collaboration has already received theoretical and practical recognition (Liu, Cheng, Yan, & Ye, 2018), and has become one of the favorite topics in the academic community of bibliometrics and scientific evaluation. Prior researches showed that the growth rate of research spanning across national borders has increased dramatically in recent years and surpassed that of domestic efforts across the globe in many scientific disciplines (Adams, 2012, 2013; Coccia & Bozeman, 2016; Hsiehchen, Espinoza, & Hsieh, 2015; Wagner, Park, & Leydesdorff, 2015), indicating the "pervasive process in motion, moving towards a truly interconnected global science system" (Waltman, Tijssen, & van Eck, 2011). Many more nations participate in global collaboration than was the case two decades ago (Bornmann & Mutz, 2015). Additionally, the number of internationally co-authored articles not only grew faster, but also cited more often than traditional "nationally-co-authored" articles (Narin, 1991; National Science Board, 2016; Persson, Glänzel, & Danell, 2004).

Conforming to the trend of the times, many countries and regions have vigorously promoted international collaborations in Big Data research and applications in recent years (Ma *et al.*, 2015). Previous research has shown that as researchers are encouraged to work and study collaboratively, and as international co-authorship has become the mainstream in Big Data research (George, Haas, & Pentland, 2014), there has been an increase in publications involving collaboration at different levels published in journals (Michael & Miller, 2013). In the context of Big Data campaign and globalization of science and technology, it is worthwhile to reveal the collaboration status worldwide, as well as its influence on academic performance in this topic. This current paper, based on bibliographic data from 2003-2017, will contribute in this regard by providing the general status of productivity and collaboration, team size in terms of number of authors, institutions and countries, as well as the relationship between collaboration characteristics and citation impact.

The remainder of this paper proceeds as follows: Section 2 introduces the data and methods used in this research. Section 3 analyzes the status of the global scientific collaboration and the relationship between collaboration performance and citation impact, in terms of different types and team sizes. Concluding remarks are provided in the last section.

Data and Methods

Data collection

To target publications related with Big Data research, existing studies mostly apply the topic retrieval strategy, using "Big Data" in the title, abstract or keywords in major academic databases (e.g., Ekbia *et al.*, 2015; Wamba *et al.*, 2015; Hu & Zhang, 2017; Peng *et al*., 2017). Drawing on these studies, we define "Big Data research" as the publications clearly identifying

the term "Big Data" and its synonyms (*i.e.* "large-scale data", "massive data" and "huge data") in our research. Publication and citation data were drawn from Web of Science (WoS).
Specifically, we retrieved publications (*i.e.* "Article", "Review" and "Proceeding papers") related to Big Data with the query: *PY=2003-2017 and TS=("big data" or "bigdata" or "huge data" or "large scale data" or "large-scale data" or "massive data")*. The period between 2003 and 2017 was selected due to the fact that research on Big Data began very recently and has gained momentum only during the last few years. Ultimately, this approach resulted in 28,201 papers. It is important to highlight here that with this approach we are adopting a narrow but precise approach: we are interested in publications explicitly using the search terms ("Big Data" and its synonyms). There is no doubt that some research may be related to Big Data research without mentioning it as such (e.g., by using other denominations – e.g., Data Science; or using Big Data related techniques, e.g., machine learning, cloud computing, etc.). However, we argue that our narrower approach is the most efficient in terms of identifying those publications that have the strongest alignment with the core concept of Big Data. The number of publications in different types is shown in table 1, among which proceeding papers account for the largest proportion (61.2%).
According to the classification of the Center for Science and Technology Studies (CWTS) in Leiden, all the journal publications (*i.e.* Articles and Reviews) can be classified into five exclusive macro-fields based on the WoS subject categories: Social Sciences and Humanities (SSH), Biomedical and Health Sciences (BHS), Physical Sciences and Engineering (PSE), Life and Earth Sciences (LES), and Mathematics and Computer Science (MCS). Considering the interdisciplinarity of Big Data research, as well as the different degree of teamwork in different fields, we will make a comprehensive analysis of the collaboration features and citation impact by fields. The numbers and proportions of these journal papers in each field are shown in Table 1. Obviously, Mathematics and Computer Science is the major field contributing the most of relevant publications in the data set, followed by Biomedical and Health Sciences and Social Sciences and Humanities, while physical sciences and engineering contributes the least.

Table 1. Numbers and proportions of journal papers in main fields

	Number of publications	Share (%)
1: Social Sciences and Humanities (SSH)	1,975	16.1
2: Biomedical and Health Sciences (BHS)	2,217	18.1
3: Physical Sciences and Engineering (PSE)	1,231	10.1
4: Life and Earth Sciences (LES)	1,489	12.2
5: Mathematics and Computer Science (MCS)	5,340	44.6
Grand Total	12,252	100.0

Indicators and variables

Collaborated papers are defined as those containing at least two authors, while those with only one author are non-collaborative papers. To be specific, when two or more institutions appear in the author's addresses of one publication, it is considered to be an institutional collaborated publication and it is counted in full for each of the contributing institutions. This classification criterion is also applicable to international collaboration papers in our research. Author names, affiliations and countries are identified by the address information.
Ordinary Least Squares (OLS) regression analysis with robust standard errors is used to test the relationship between collaboration features and citation impact. This is carried out in two steps: first, the association between collaboration types and citation impact is tested; next, the relationship between team sizes (measured as number of authors per institution, number of

institutions per country, number of countries per article) and citation impact is explored. Variables are described in the following section.

Dependent variable is citation impact, measured as mean normalized citation score [MNCS]. This is one of the most frequently used field-normalized indicators and is basically calculated by dividing the citation count of a focal paper by the average citation count of the papers published in the same field and publication year (Bozeman, Dietz, & Gaughan, 2001).

Two types of independent variables are used. One is collaboration type, involving four types according to the number of participants:

- Non-collaborative publications [NOC]: documents with only one author in the address;
- Domestic collaborative publications [DC]: documents with at least two authors in one country, and it can be divided into two types as follows;
    - Within-institution collaborative publications [WIC]: documents with at least two authors in the same institution;
    - Cross-institution collaborative publications [CIC]: documents with at least two authors and institution addresses in the same country;
- International collaborative publications [IC]: documents with at least two authors and institution addresses from different country.

Collaborative papers are classified according to collaborative type, following the ‘‘maximum distance principle’’. If a paper belongs to more than one collaborative type, the priority rule will be shown as IC>DC (CIC>WIC)>NOC.

Another independent variable is team size at different levels:

- Number of authors [N_au]: number of authors that signed a document;
- Number of institutions [N_ins]: number of distinct institutions to which authors of the document were affiliated;
- Number of countries [N_cu]: number of distinct countries found in the authors' addresses.

Variables related to document characteristics are controlled in the regression analysis.

- Publication year [PY]: the year when the publication is published;
- Document type [DT]: Article [A], Review [R] or Proceeding papers [P];
- Number of references [NR]: number of cited references in the reference list. To the extent that the references influence the paper by providing information and ideas from other scientists, they can be viewed as indicators of “invisible coauthors,” self-citations aside (Freeman, Ganguli, & Murciano-Goroff, 2014).

Results

General profile of publications and collaboration

Figure 2 shows the annual number of Big Data publications and the percentages of publications in the five fields for each year. A steady increase in total output of Big Data related research from 2003 to 2017 can be observed, especially since the year 2012 with an exponential growth ($R^2 > 0.90$). In 2017, the global output has reached over 7000, involving more than 100 countries worldwide, of which developed countries and some emerging economics are the main contributors. Among all the countries, China and USA are the most prolific in the whole period, far exceeding the others, and their status as science superpowers is ongoing. As for the research fields, although Mathematics and Computer Science occupies the most (over 44%), it has decreased slightly over time, while a significant increase in Social Sciences and Humanities is clearly observed. The other three fields seem to be relatively stable in proportion.

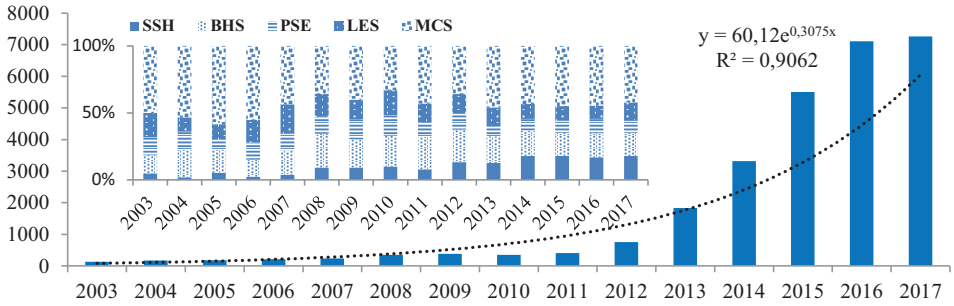


Figure 2. Global annual output and percentages of publications in five fields.

Globally, domestic collaboration is the main type for Big Data research (over 70%). Within-institutional collaboration remains above 40%, despite of a slight decrease over time. Relatively speaking, international collaboration occurs few in Big Data research, accounting for only about 16%. However, its proportion gradually grew to 24% in 2017 with the development of technology globalization, almost four times that in 2003 (about 6%). Meanwhile, most (about 77%) of the international collaboration is bilateral. The share of non-collaborated papers is relatively stable, remaining at around 10%. The lion's share of co-authored papers shows a clear trend towards more active and closer scientific collaboration in Big Data research. (Figure 3)

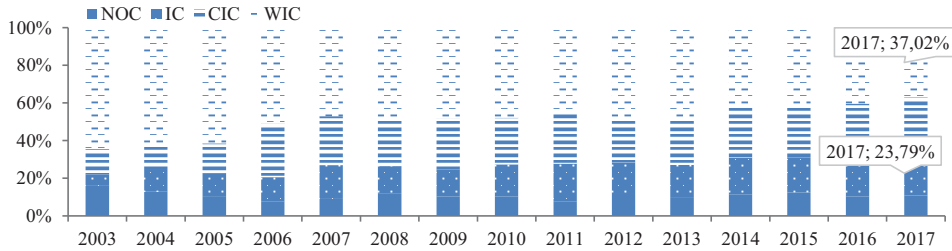


Figure 3. Percentages of papers of different collaborative types

In terms of the publications in fields, the annual percentages of different collaboration types are shown in Figure 4. Obviously, there is a tendency in all the five fields to participate more in cross-institutional or international collaboration, and the proportions of within-institutional collaboration have decreased. Specifically, for Mathematics and Computer Science, as well as Social Sciences and Humanities, international co-authorship has been the main collaboration type accounting for the largest and up-going proportion. However, domestic co-authorship, especially the collaboration among institutions is the mainstream for researchers fall within Biomedical and Health Sciences, Physical Sciences and Engineering and Life and Earth Sciences, among which Physical Sciences and Engineering shows the highest propensity to work with colleagues from other institutions. Moreover, the fairly low shares of within-institutional co-authorship in Biomedical and Health Sciences, and Mathematics and Computer Science, illustrate a high degree of collaboration among authors in these two fields.

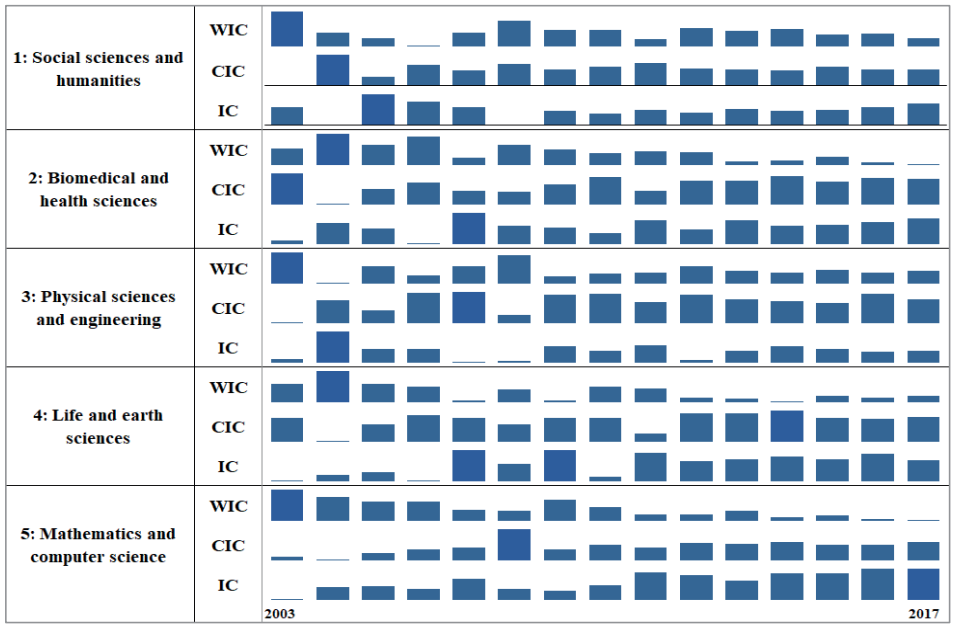


Figure 4. Annual percentages of different collaboration types in fields.

Team size of different collaboration types

Figure 5 presents the team size of different collaboration types in terms of number of authors, institutions and countries per paper. All items show the overall average of all papers for reference. It is quite reasonable that the higher the collaborative level is, the more the co-authors and institutions will be. To be specific, WIC shows a smaller number of authors per paper on average (around 3.46) than IC and CIC, while IC has much more authors (5.81 on average). As mentioned above, the lower degree of IC also can be reflected in the average number of countries (less than 3), indicating a limited number of collaborative countries on Big Data research.

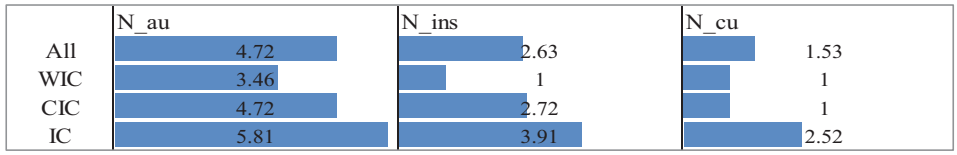


Figure 5. Distribution of different collaboration types

Team size differs considerably from one field to another. Biomedical and Health Sciences show the largest team size of about 5 authors, 3 institutions and 2 countries on average, indicating a higher degree of collaboration, followed by Physical Sciences and Engineering. In contrast, Social Sciences and Humanities is the field with the minimal size of researchers, far below the average level. With time going on, team size has increased or been stable in general, except for a slight decrease in number of institutions in Physical Sciences and Engineering (PSE), and, in number of countries in Social Sciences and Humanities (SSH) and Biomedical and Health Sciences (BHS). According to ANOVA tests, a significant increase can be observed in number of participants in Mathematics and Computer Science (MCS), and Biomedical and Health Sciences (BHS), but no statistical difference in the other fields. (Figure 6)

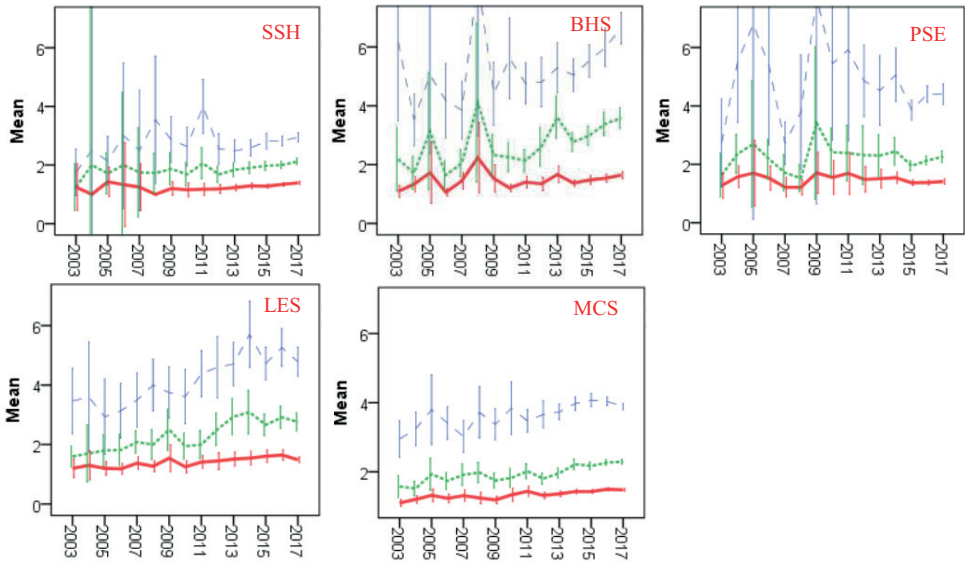


Figure 6. Team size of different fields by year. Dashed, dotted and solid lines correspond to average number of authors, institutions and countries, respectively.

Collaboration and citation impact

Figure 7 presents the annual distribution of MNCS received by papers of different types of collaboration. In general, co-authored papers (denoted as "C" in the figure) usually outnumbers non-coauthored papers in citation. In fact, despite the high values that appeared in earlier years, NOC always ranks the bottom, indicating the citation advantage of collaboration. When it comes to the three types of collaboration, it can be observed that IC usually tops the list, followed by CIC and WIC in order. Moreover, with time goes by, the citation trends of collaboration types tend to be more stable, so are the differences among them. To be specific, the citation advantage of IC is more prominent than the other two, but for WIC, the spillover effect of CIC is not that obvious. Such distribution stands to reason that collaborated papers, especially the IC papers, generally receive the most citations, followed by the CIC papers, and those with single author gain the least.

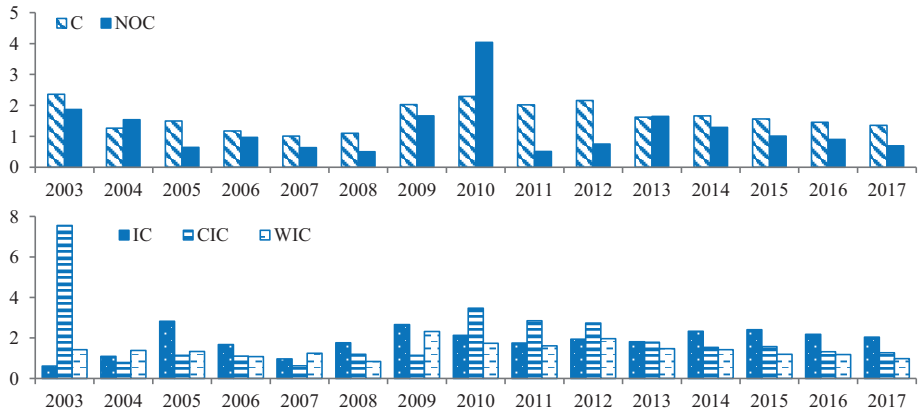


Figure 7. Annual MNCS of different collaboration types on average.

Collaboration type is confirmed to be significantly associated with citation impact measured by MNCS, as shown in Table 2. Holding other factors constant (DT, PY, NR), NOC, WIC and CIC tend to exhibit lower MNCS reflected by negative and significant coefficients of NOC, WIC, and CIC. This confirms the citation advantage of international collaboration to non- or domestic collaboration, indicating that collaborating with foreign researchers is beneficial to promote the scientific influence of a nation in general. Among domestic publications, NOC is inferior to domestic collaborated publications (WIC and CIC), and WIC shows lower citation impact than CIC. Moreover, the result is robust since same results can be found when dependent variable is set as top-1% highly-cited paper or not (0, 1) using logistic regression analysis.

Table 2. OLS Regression model for collaboration types and MNCS

ln (MNCS)	Coef.	Robust Std. Err.	t/F	p-value
NOC [a]	-212	.0154	-13.74	0.000
WIC	-.158	.0123	-12.87	0.000
CIC	-.134	.0130	-10.32	0.000
IC	--			
$Coef_{NOC}=Coef_{CIC}$ [b]			32.85	0.000
$Coef_{WIC}=Coef_{CIC}$			5.80	0.016
$Coef_{NOC}=Coef_{WIC}$			18.39	0.000
DT(P)	--			
DT(A)	.111	.0108	10.37	0.000
DT(R)	-.136	.0239	-5.69	0.000
Ln (NR)	.226	.0065	34.69	0.000
Year	fixed			
_cons	-.274	.0219	-12.51	0.000
N	27570			
F	206.27			
R^2	.1219			

a: Collaboration type is a categorical variable, therefore, positive or negative coefficients of NOC, WIC, CIC means relatively higher or lower citation impact compared with IC. b: The difference of among NOC, WIC, and CIC is tested via F test, for example, $p(Coef_{WIC}=Coef_{CIC})=0.016$ indicates significant difference between the coefficients of WIC and CIC, consequently, the WIC exhibits a relatively lower citation impact compared with CIC.

Taking research fields into account, results in Table 3 further confirm the excellence of international collaborated publications in citation impact in all fields, as shown by negative and significant coefficients of NOC, WIC and CIC. Citation variation among NOC, WIC and CIC differs in fields: collaboration helps to improve citation impact in all the five fields, especially in Physical Sciences and Engineering (PSE) and Life and earth sciences (LES), citation impact of WIC and CIC shows no statistical difference, and are both higher than that of NOC; for Biomedical and Health Sciences (BHS), only difference between NOC and CIC's citation impact is significant; while for the other two. For most of the fields (*i.e.* Social Sciences and Humanities, and Mathematics and Computer Science), no difference exists in citation impact of NOC, WIC and CIC.

The relationship between number of participants and citation impact within different types of collaboration are reported in Table 4. First, number of authors is positively related to citation impact: the coefficients translates into 0.075%, 0.037% and 0.07% higher MNCS when the number of authors increases by 1% in WIC, CIC and IC respectively. Second, more institutions in CIC doesn't generate any effect, whereas more institutions involved in IC has a positive association with citation impact, *i.e.*, 0.069% higher MNCS is associated with 1% more of number of institutions in IC. Third, 1% more countries in IC results in 0.204% more MNCS, that is, publications with authors from multiple nations are of higher impact than domestic

publications in terms of citations, and more countries in international collaboration relates to higher impact. To conclude, more participants at different level in a paper is related to higher citation impact in general, except that more institutions in CIC does not generate extra citation increment.

Table 3. OLS Regression model for collaboration types and MNCS in fields

	SSH	BHS	PSE	LES	MCS
NOC	**-0.189*****	**-0.162*****	**-0.267*****	**-0.176*****	**-0.212*****
	(0.050)	(0.049)	(0.062)	(0.049)	(0.040)
WIC	**-0.129****	**-0.109*****	**-0.102****	**-0.072***	**-0.225*****
	(0.054)	(0.035)	(0.041)	(0.041)	(0.023)
CIC	**-0.111****	**-0.058***	**-0.113****	**-0.075****	**-0.188*****
	(0.050)	(0.032)	(0.044)	(0.038)	(0.024)
$F(Coef_{NOC}=Coef_{CIC})$	2.55	**4.68****	**6.61****	**4.29****	0.38
$F(Coef_{WIC}=Coef_{CIC})$	0.11	2.23	0.08	0.01	0.11
$F(Coef_{NOC}=Coef_{WIC})$	1.31	1.19	**8.13*****	**4.74****	0.11
Observations	1912	2199	1207	1483	5273
F	16.467	7.563	9.379	15.452	28.833
R^2	0.110	0.060	0.112	0.148	0.115

Robust standard errors in parentheses: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Controlled variables are omitted.

Table 4. OLS Regression model for team sizes and MNCS

ln (MNCS)	WIC	CIC	IC
ln (# authors)	**0.075*****	**0.037***	**0.070*****
	(0.016)	(0.020)	(0.026)
ln (# institutions)	.	0.043	**0.069****
	.	(0.030)	(0.034)
ln (# countries)	.	.	**0.204*****
	.	.	(0.051)
DT (A)	0.086***	0.093***	0.077***
	(0.016)	(0.020)	(0.027)
DT (R)	-0.105**	-0.133***	-0.250***
	(0.041)	(0.047)	(0.052)
log (NR)	0.211***	0.226***	0.258***
	(0.010)	(0.013)	(0.018)
Year fixed	yes	yes	yes
cons	-0.480***	-0.483***	-0.706***
	(0.032)	(0.045)	(0.064)
N	11620	7525	5647
F	71.187	44.620	33.567
R^2	0.091	0.100	0.109

Robust standard errors in parentheses: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Controlled variables are omitted.

The contribution of team size varies a lot in journal articles of different fields as shown in Table 5. Relationships between team size and MNCS in Physical Sciences and Engineering (PSE) are similar to that in all the publications. Cases in WIC and CIC of Life and Earth Sciences (LES) are also similar, while for IC in LSE, only number of countries counts. In the other fields (*i.e.* SSH, BHS, and MCS), only part of the significant predictors in models in Table 4 are statistically related to citation impact. Such contrast indicates that relationship between team sizes and citation impact differs among fields.

Table 5. OLS Regression model for team sizes and MNCS in fields

	Ln (MNCS)	SSH	BHS	PSE	LES	MCS
WIC	ln (# authors)	0.097	**0.103****	-0.077	**0.131****	**0.063***
		(0.113)	(0.050)	(0.063)	(0.066)	(0.037)
	Observations	402	553	416	394	1721
	F	.	3.086	6.072	4.644	9.689
	r^2	0.104	0.091	0.151	0.165	0.095
CIC	ln (# authors)	-0.041	0.027	**0.143****	**0.138***	**0.078***
		(0.091)	(0.049)	(0.068)	(0.075)	(0.043)
	ln (# institutions)	**0.359*****	0.040	-0.088	-0.009	0.053
		(0.134)	(0.060)	(0.120)	(0.099)	(0.066)
	Observations	530	845	366	500	1532
	F	.	4.155	.	.	7.326
	R^2	0.138	0.055	0.134	0.137	0.094
IC	ln (# authors)	0.176	**0.130****	0.036	0.022	**0.126*****
		(0.118)	(0.057)	(0.071)	(0.067)	(0.047)
	ln (# institutions)	0.045	0.086	-0.050	0.048	**0.102***
		(0.140)	(0.070)	(0.100)	(0.093)	(0.059)
	ln (# countries)	**0.401***	0.059	0.224	**0.251****	**0.281*****
		(0.241)	(0.086)	(0.150)	(0.114)	(0.101)
	Observations	465	635	344	450	1724
	F	.	.	3.217	4.632	21.526
	R^2	0.138	0.127	0.086	0.176	0.106

Robust standard errors in parentheses: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Controlled variables are omitted.

Discussion and Conclusion

Scientific output of Big Data related research shows a steady increase during 2003-2017, especially since the year 2012. China and USA are the most prolific ones. Collaboration happens mainly within a nation, more specifically, within an institution. Such a phenomenon may be caused by personal mobility and the national funding policies of large countries (Larivière, Vincent, & Sugimoto *et al.*, 2009). International collaboration grows faster despite of its small share. Among all the research fields involved, Mathematics and Computer Science is the main contributor to Big Data publications, which also has the highest and growing proportion of international co-authored papers.

Because of the nature of research and the way the research work be organized, field-oriented otherness in collaboration is a common rule and a self-evident feature in science development (Puuska, Muhonen, & Leino, 2014). To be precise, Biomedical and Health Sciences shows the highest degree of collaboration, followed sequentially by Physical Sciences and Engineering, Life and Earth Sciences, as well as Mathematics and Computer Science. In comparison, Big Data research in Social Sciences and Humanities is as expected the least collaborated. Although the degree of collaboration differs among fields, the growth trend of proportion of collaborated publications is common, along with the expansion of tame size, especially measured by the number of authors.

Last but not least, evidence is presented that collaboration is statistically associated with higher citation impact, but such association depends on the types of collaboration and team sizes, and, varies among different fields. Collaborating with authors from other institutions, especially from foreign institutions generates extra citation increment; in journal articles of 5 fields, citation variation among different domestic collaboration (NOC, WIC, and CIC) might be insignificant in some fields. As for the number of participants within different types of collaborated publications, a larger size of a team correlates with higher citation impact.

However, such citation benefit also varies among fields, with indicators of team size serve as significant predictors of citation impact in various collaboration types of different fields, of which Mathematics and Computer Science shows the most significant impact by both indicators, especially the types, while Physic and Engineering Science the least.
Our results suggest that, in general, enlarging collaboration size can be an option to receive more citation impact for researchers from all fields. With the same team size, collaborating with international partners can be the first choice. When international collaboration is not possible, domestic and cross-institutional collaboration can be an alternative.

Acknowledgments

This work is funded by National Natural Science Foundation of China (grant number: 71473219, 71843012).

Notes

1. Source: https://www.statista.com/statistics/254266/global-big-data-market-forecast/

References

Adams, J. (2012). Collaborations: The rise of research networks. *Nature, 490*(7420), 335–336.
Adams, J. (2013). Collaborations: The fourth age of research. *Nature, 497*(7451), 557–560.
Bornmann, L., & Mutz, R. (2015). Growth rates of modern science: A bibliometric analysis based on the number of publications and cited references. *Journal of the Association for Information Science and Technology, 66*(11). https://doi.org/10.1002/asi.23329
Bozeman, B., Dietz, J. S., & Gaughan, M. (2001). Scientific and technical human capital: an alternative model for research evaluation. *International Journal of Technology Management, 22*(7/8), 716. https://doi.org/10.1504/IJTM.2001.002988
Chen, C. L., & Zhang, C. (2014). Data-intensive applications, challenges, techniques and technologies: A survey on Big Data. *Information Science, 275*, 314-347.
Clarke, L., & Plume, A. (2011). *Knowledge, networks and nations: Global scientific collaboration in the 21st century*. London: The Royal Society.
Coccia, M., & Bozeman, B. (2016). Allometric models to measure and analyze the evolution of international research collaboration. *Scientometrics, 108*(3), 1065–1084.
Ekbia, H., Bowman, T., & Weingart, S. (2015). Big Data, Bigger Dilemmas: A Critical Review. *Journal of the Association for Information Science and Technology, 66*(8), 1523–1545.
European Commission. (2003). The Third European Report on Science and Technology Indicators 2003. *Towards a knowledge-based economy*. Brussels: European Commission.
Freeman, R. B., Ganguli, I., & Murciano-Goroff, R. (2014). Why and wherefore of increased scientific collaboration. *National Bureau of Economic Research*.
George, G., Haas, M. R., & Pentland, A. (2014) Big Data and management. *Academy of Management Journal 57*(2): 321–326.
Haustein, S., Costas, R., & Larivière, V. (2015). Characterizing social media metrics of scholarly papers: The effect of document properties and collaboration patterns. *PLoS ONE, 10*(3), 1–21.
Hsiehchen, D., Espinoza, M., & Hsieh, A. (2015). Multinational teams and diseconomies of scale in collaborative research. *Science Advances, 1*(8), e1500211. https://doi.org/10.1126/sciadv.1500211
Hsiehchen, D., Espinoza, M., & Hsieh, A. (2018). Evolution of collaboration and optimization of impact: self-organization in multinational research. *Scientometrics, 117*(1), 391–407.
Huang, Y., Zhang, Y., Youtie, J., Porter, A. L., & Wang, X. (2016). How does national scientific funding support emerging interdisciplinary research: A comparison study of Big Data research in the US and China. *PLoS ONE, 11*(5), e0154509. https://doi.org/10.1371/journal.pone.0154509
Kitchin, R. *The Data Revolution: Big Data, Open Data, Data Infrastructures and Their Consequences*. Thousand Oaks, CA: Sage Publishing, 2014.
Larivière, V., Gingras, Y., Sugimoto, C. R., & Tsou, A. (2009). Team Size Matters: Collaboration and Scientific Impact Since 1900. *Communications in Information Literacy, 3*(2), 80–90.

Liu, Y., Cheng, Y., Yan, Z., & Ye, X. (2018). Multilevel Analysis of International Scientific Collaboration Network in the Influenza Virus Vaccine Field: 2006–2013. *Sustainability, 10*(4), 1232. https://doi.org/10.3390/su10041232

Ma, H., Fang, C., Pang, B. & Wang, S. (2015). Structure of Chinese city network as driven by technological knowledge flows. *Chinese Geographical Science, 25*(4), 498-510.

Michael, K., & Miller, K. W. (2013). Big Data: New Opportunities and New Challenges. *IEEE Computer Society*, *46*(6), 22–24. https://doi.org/10.1109/MC.2013.196

Narin, F. (1991). Globalisation of research, scholarly information and patents—ten year trends. *The Serials Librarian*, *21*(2-3), 33-44.

National Science Board. (2016). *Science and Engineering Indicators 2016*. https://doi.org/10.1002/ejoc.201200111

Peng, Y., Shi, J., Fantinato, M., & Chen, J. (2017). A study on the author collaboration network in Big Data*. *Information Systems Frontiers, 19*(6), 1329–1342.

Persson, O., Glänzel, W., & Danell, R. (2004). Inflationary bibliometric values: the role of scientific collaboration and the need for relative indicators in evaluative studies. *Scientometrics, 60*(3), 421-432.

Puuska, H. M., Muhonen, R., & Leino, Y. (2014). International and domestic co-publishing and their citation impact in different disciplines. *Scientometrics, 98*(2), 823–839.

Richard B. Freeman, Ina Ganguli, R. M.-G. (2014). *Why and wherefore of increased scientific collaboration*. National Bureau of Economic Research. http://doi.org/10.1007/s13398-014-0173-7.2

Van Eck, N. J., & Waltman, L. (2010). Software survey: VOSviewer, a computer program for bibliometric mapping. *Scientometrics, 84*(2), 523-538.

Wagner, C. S. (2008). *The new invisible college: Science for development*. Washington: Brookings Institution Press.

Wagner, C. S., & Jonkers, K. (2017). Open countries have strong science. *Nature, 550*(7674), 33.

Wagner, C. S., Park, H. W., & Leydesdorff, L. (2015). The continuing growth of global cooperation networks in research: A conundrum for national governments. *PLoS ONE, 10*(7).

Waltman, L., Tijssen, R. J. W., & van Eck, N. J. (2011). Globalisation of science in kilometres. *Journal of Informetrics, 5*(4), 574–582. Digital Libraries.

Wamba, S. F., Akter, S., Edwards, A., F, G. C., & Gnanzou, D. (2015). How 'Big Data' can make big impact: Findings from a systematic review and a longitudinal case study Samuel. Intern. *Journal of Production Economics, 165*, 234–246.

Examining the citation and altmetric advantage of bioRxiv preprints

Nicholas Fraser[1*], Fakhri Momeni[2], Philipp Mayr[2] and Isabella Peters[1,3]

*n.fraser@zbw.eu
[1] ZBW – Leibniz Information Centre for Economics, Kiel, Germany
[2] GESIS – Leibniz Institute for the Social Sciences, Cologne, Germany
[3] Kiel University, Kiel, Germany

Abstract

Early dissemination of scientific results in the form of preprints is an important component of modern open science workflows. A potential motivation for scientists to deposit preprints is to enhance the citation and/or social impact of their work, an effect which has been empirically observed for preprints deposited to arXiv, a preprint server primarily for research in physics, astronomy and mathematics. Here we report work in progress on a study investigating the extensibility of these findings to the biological sciences, by assessing the citation and altmetric advantage of depositing preprints to the preprint server bioRxiv. We retrieved article metadata together with citation and altmetric counts for a cohort of >8000 articles that were deposited to bioRxiv as preprints, and compare them with a control group of non-deposited articles. We find that citation and altmetric counts (tweets and Mendeley reads) are higher for articles that were deposited to bioRxiv than those that were not. Future work will aim to statistically quantify the effect of multiple confounding variables on this relationship, and to investigate which features of papers or authors may drive the preprint citation and altmetric advantage of bioRxiv.

Introduction

Preprints, typically defined as versions of scientific articles that have not yet been formally accepted for publication in a peer-reviewed journal, are an important feature of modern scholarly communication (Berg et al., 2016). Major motivations for the scholarly community to adopt the use of preprints have been proposed as *early discovery* (manuscripts are available to the scientific community earlier, bypassing the time-consuming peer review process), *open access* (manuscripts are publicly available without having to pay expensive fees or subscriptions) and *early feedback* (authors can receive immediate feedback from the scientific community to include in revised versions) (Maggio et al., 2018). An additional motivation for scholars to deposit preprints may be to increase citation counts and/or altmetric indicators such as shares on social media platforms. For example, recent surveys conducted by the Association for Computational Linguistics (ACL) and Special Interest Group on Information Retrieval (SIGIR) found that 32 and 15 % of respondents were respectively motivated to deposit preprints "to maximize the paper's citation count" (Foster et al., 2017; Kelly, 2018).

A body of evidence has emerged which supports the notion of a citation differential between journal articles that were previously deposited as preprints and those that were not, with several studies concluding that arXiv-deposited articles subsequently received more citations than non-deposited articles (Davis and Fromerth, 2007, Moed, 2007; Gentil-Beccot et al., 2010; Larivière et al., 2014). Multiple factors have been proposed as drivers of this citation differential, including increased readership due to wider accessibility (the "open access effect"), earlier accumulation of citations due to the earlier availability of articles to be read and cited (the "early access effect"), authors preferential deposition of their highest quality articles as preprints (the "self-selection effect"), or a combination thereof (Kurtz et al., 2005). Whilst citation differentials have been well documented for articles deposited to arXiv, the long-established nature of depositing preprints in physics, astronomy and mathematics may make it unsuitable to extend the conclusions of these studies to other subject-specific preprint repositories, where preprint deposition is less established.

bioRxiv is a preprint repository aimed at researchers in the biological sciences, launched in November 2013 and hosted by the Cold Spring Harbor Laboratory (https://www.biorxiv.org/). As a relatively new service, it presents an interesting target for

analysing impact metrics in a community where preprints have been under-utilised in comparison to the fields of physics, astronomy and mathematics (Ginsparg, 2016). A previous study by Serghiou and Ioannidis (2018) provided insights into the potential citation and altmetrics advantage of bioRxiv-deposited article over non-deposited articles, finding that bioRxiv-deposited articles had significantly higher citation counts and altmetric scores than non-deposited articles. However, this study was based on a relatively small sample of 776 preprints that could be matched to published articles, and did not attempt to provide any statistical constraints on the potential effect of confounding variables, nor assess longitudinal trends in citations or altmetric indicators.

In the following "Work in Progress" paper we present methods and initial results of an investigation into the potential citation and altmetric advantage for a large amount of articles that have been deposited as bioRxiv preprints. In our future work, we shall aim to provide statistical constraints on the size of this effect and the role of multiple confounding variables to complement and build upon the initial work by Serghiou and Ioannidis (2018).

Methods

Preprint and Article Metadata

Metadata of all preprints submitted to bioRxiv between November 2013 and December 2017 were harvested via the Crossref public Application Programming Interface (API) (N = 18,839). Of these preprint records, 9,191 included 'relationship' properties which provide a DOI link to the respective final published version of an article, typically in a journal (herein referred to as 'bioRxiv-deposited articles'). These links are maintained and routinely updated by bioRxiv through monitoring of databases such as Crossref and PubMed, or through information provided directly by the authors (personal correspondence with bioRxiv representative). Each DOI contained in the 'relationship' property was queried via the Crossref API to retrieve the metadata record of the published article. Nine duplicate records were identified, likely caused by authors uploading multiple preprint versions as individual records. In these cases, we retained the earlier posted record in our sample and discarded the later record. A further six records were found to contain incorrect DOI information for the published article (Crossref API resolved to a 404 error), and were also discarded.

Crossref records of published articles were matched to records in Clarivate Analytics Web of Science (WoS) (leveraging the data infrastructure of the German Competence Centre for Bibliometrics: http://www.forschungsinfo.de/Bibliometrie/en/index.php) via direct, case-insensitive correspondence between DOIs or titles. WoS records were limited to 'journal' publication types, 'article' or 'review' document types, and records with reference counts greater than zero, to reduce the rare incidence of editorial material incorrectly classified as 'article' type documents. For 12 articles, duplicate records were identified in WoS and were therefore discarded. Following these steps, 6,812 Crossref records (74 %) were successfully matched to a WoS record. The reason for the relatively low percentage of matches is that a large proportion of preprints in our dataset were deposited in mid-late 2017 and thus the final journal articles were only published in 2018. To promote reproducibility, we use a WoS database 'snapshot' which only partially covers 2018 and thus many of the later publications are missed. However, for publication years 2013 to 2017 we are able to match >90 % of Crossref records to a WoS record.

A manual Google search of a small sample of preprints that did not have a Crossref 'relationship' property revealed that a significant percentage were in fact published in another format (journal article, book chapter, conference paper) subsequent to their deposition on bioRxiv, but not linked via Crossref. To partially account for these missing links, we performed an additional matching procedure between bioRxiv preprints and WoS records

(limited to those not matched in the previous step), based upon direct correspondence between the last name of the first author and fuzzy matching of the article title OR first 100 characters of the abstract for the bioRxiv preprint and WoS record. Fuzzy matching was conducted with the R package 'stringdist' (van der Loo, 2018), using the Jaro-Winckler distance algorithm and a similarity of 80 %. Matches were further validated by comparison of the author count of the preprint record and WoS record. This resulted in retrieval of WoS records of a further 1,476 bioRxiv-deposited articles, which were merged with the previous set to create a full set of 8,288 bioRxiv-deposited articles.

Control Group

To conduct comparative analysis between bioRxiv-deposited articles (as defined in the previous section) and non-deposited articles, it is necessary to generate a control group of non-deposited articles. As a first step we retrieved from WoS all articles published in the same journal-issues as the articles within our dataset of bioRxiv-deposited articles, limited to 'journal' publication types, 'article' or 'review' document types, and records with reference counts greater than zero. Articles present in the bioRxiv-deposited group were removed from the control group.

A matching process was then conducted to match each bioRxiv-deposited article with a single, random article published in the same journal-issue in the control group. A potential weakness of this matching procedure lies in the inclusion of articles published within multidisciplinary journals (e.g. PLOS One, Scientific Reports), as it would be unwise to match a biology-focused article with an article from another discipline with drastically different publication and citing behaviour. For articles published in multidisciplinary journals, we therefore conducted an additional procedure in which articles in both the bioRxiv-deposited and non-deposited groups were re-classified into WoS categories based on the most frequently cited categories amongst their references (modified from the multidisciplinary article classification procedure of Piwowar et al., 2018). Where categories were cited equally frequently, articles were assigned to multiple categories. For each bioRxiv-deposited article, a single, random non-deposited article was selected from the same journal-issue and categories in the control group. In total, 8,194 articles from the set of 8,288 bioRxiv-deposited articles could be matched with a non-deposited control article – the remainder could not be matched (e.g. when no other non-deposited articles were published in the journal in the same month) and were discarded from our analysis.

Publication Dates

A metholodogical consideration when analysing citation data is in the treatment of publication dates. Publication dates for individual articles are reported by multiple outlets (e.g. by Crossref, WoS and the publishers themselves), but often represent different publication points, such as the date of DOI registration, the WoS indexing date, or the online and print publication dates reported by the publisher (see Haustein et al., 2015, for a discussion on the lack of standardization and difficulty in reconciling publication dates from multiple sources). In our study, we implement the Crossref 'created-date' property as the canonical date of publication for all articles and citing articles in our datasets, in line with the approach of Fang and Costas (2015). The 'created-date' is the date upon which the DOI is first registered and can thus be considered a good proxy for the first online availability of an article at the publisher website. An advantage of this method is that we can report citation counts at a monthly resolution, as recently advocated by Donner (2018), which may be more suitable than report annually-resolved citation counts due to the relatively short time-span of our analysis period and rapid growth of bioRxiv.

Citation Data

Metadata of citing articles were retrieved from WoS for all bioRxiv-deposited and non-deposited articles, and citing article DOIs subsequently queried against the Crossref API to retrieve publication dates. In total we retrieved records of 49,368 articles citing bioRxiv-deposited articles, and 35,389 articles citing non-deposited articles. Citation counts were aggregated at a monthly level for each article. As citation counts typically exhibit a Log-Normal distribution (Ruocco et al., 2017), we additionally log-transformed all aggregated citation counts prior to reporting.

Altmetrics Data

Altmetric data, including tweets and Mendeley reads, were retrieved for all bioRxiv-deposited and non-deposited articles by querying their DOIs against the Altmetric.com API (https://api.altmetric.com/).

Results

Development of bioRxiv preprints

Since launching in November 2013, bioRxiv has grown rapidly in terms of preprint deposits (Figure 1). One and two year probabilities of journal article publication are found to be 55.6 % and 64.9 %, respectively, slightly higher than estimates of 48.0 % and 55.5 % of Serghiou and Ioannidis (2018), likely due to our improved preprint-article matching procedure. The median review time for bioRxiv preprints is found to be 157 days, in comparison to a field-wide average of approximately 100 days in biomedical sciences (Powell, 2016). Discrepancies between these timescales can be attributed to multiple factors, including: (1) authors may not submit their preprint to a journal immediately following the deposit of their preprint, (2) a preprint may be rejected by one or more journals prior to acceptance, thus our time difference represents multiple review cycles, and (3) bioRxiv preprints may be preferentially submitted to journals with longer than average review times.

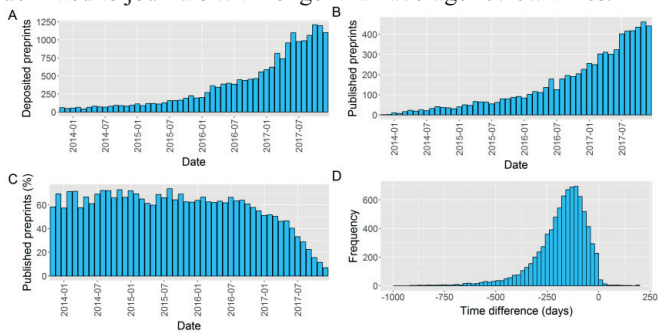


Figure 1: (A) distribution of bioRxiv preprint deposits; (B) Monthly distribution of published papers with bioRxiv preprints; (C) Monthly percentage of bioRxiv preprints that have been subsequently published (month refers to preprint deposit date); (D) Frequency distribution of days between publication date and preprint deposit of journal article.

Citations Analysis

Monthly average citations rates for bioRxiv-deposited and non-deposited articles are shown in Figure 2. We limited our dataset to articles and citing articles published between November 2013 and December 2017, and limit our results to a 36-month citation period due to the low numbers of articles with longer citation histories available. Figure 2 shows a clear divergence between the two groups, with bioRxiv-deposited articles being cited more frequently than non-deposited articles in the same months.

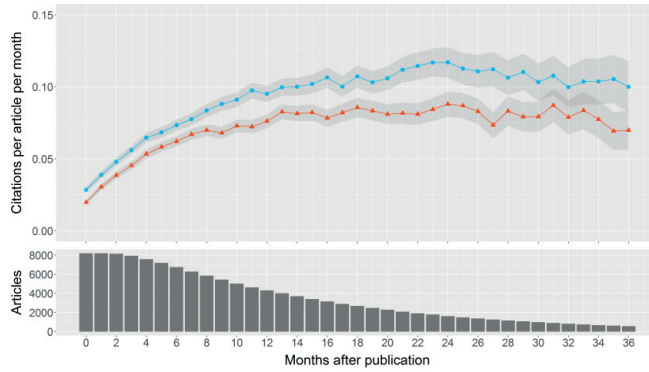


Figure 2: Upper panel: average citations per article per month (log-transformed) of bioRxiv-deposited articles (blue circles) and non-deposited articles (red triangles), grey shading represents 95 % confidence interval. Lower panel: number of articles included at each time step.

Altmetrics Analysis

Distributions of tweets and Mendeley reads for bioRxiv-deposited and non-deposited articles are shown in Figure 3. Wilcoxon signed-rank tests (a non-parametric test for comparing distributions of two matched samples; Wilcoxon (1945)) were conducted to compare altmetric indicators between groups, and found that altmetric values were statistically significantly higher in the bioRxiv-deposited articles compared to the non-deposited articles for tweets ($Z = -21.25$, $p < 0.001$, $r = 0.23$) and Mendeley reads ($Z = -17.42$, $p < 0.001$, $r = 0.19$).

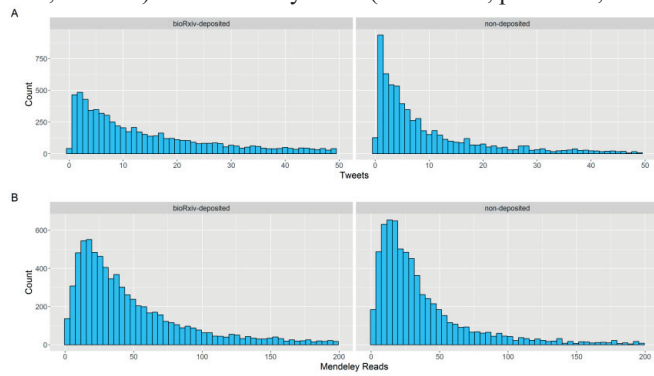


Figure 3: Frequency distribution of altmetric counts for bioRxiv-deposited (left) and non-deposited articles (right), including (A) number of tweets, and (B) number of Mendeley reads.

Future Work

Future work in this study will focus on deeper statistical analyses to quantify the citation and altmetric advantage of bioRxiv preprints at key time periods (e.g. at 12 months, 24 months and 36 months post-publication), and to account for the potential role of confounding variables on this advantage (e.g. article open access status, journal impact factor, article age, author count, author countries, author seniority, author gender, etc). We will also aim to expand on some aspects of our methodology, for example incorporating a wider variety of altmetrics indicators, and by expanding our citation analysis to consider the volume and role of citations made directly to preprints themselves. We will frame our results within the context of previous studies which have studied the citation advantage of preprints on arXiv – for example, do our results show an *early access* effect, in which citations are accelerated by

bioRxiv (as found in the arXiv context by Moed (2007)), or can they be better explained by an *open access* or *self selection* effect?

Acknowledgements

This work is supported by BMBF project OASE, grant number 01PU17005A.

References

Berg, J. M., Bhalla, N., Bourne, P. E., Chalfie, M., Drubin, D. G., Fraser, J. S., … Wolberger, C. (2016). Preprints for the life sciences. Science, 352(6288), 899–901. https://doi.org/10.1126/science.aaf9133

Davis, P. M., & Fromerth, M. J. (2007). Does the arXiv lead to higher citations and reduced publisher downloads for mathematics articles? Scientometrics, 71(2), 203–215. https://doi.org/10.1007/s11192-007-1661-8

Donner, P. (2018). Effect of publication month on citation impact. Journal of Informetrics, 12(1), 330–343. https://doi.org/10.1016/j.joi.2018.01.012

Fang, Z., & Costas, R. (2018). Studying the posts accumulation patterns of Altmetric.com data sources. Presented at the Altmetrics18 Conference.

Foster, J., Hearst, M., Nivre, J., & Zhao, S. (2017). Report on ACL Survey on Preprint Publishing and Reviewing. Association for Computational Linguistics.

Gentil-Beccot, A., Mele, S., & Brooks, T. C. (2010). Citing and reading behaviours in high-energy physics. Scientometrics, 84(2), 345–355. https://doi.org/10.1007/s11192-009-0111-1

Ginsparg, P. (2016). Preprint Déjà Vu. The EMBO Journal, 35(24), 2620–2625. https://doi.org/10.15252/embj.201695531

Haustein, S., Bowman, T. D., & Costas, R. (2015). When is an article actually published? An analysis of online availability, publication, and indexation dates. ArXiv: 1505.00796.

Kelly, D. (2018). SIGIR Community Survey on Preprint Services. ACM SIGIR Forum, 52(1), 11–33. https://doi.org/10.1145/3274784.3274787

Kurtz, M. J., Eichhorn, G., Accomazzi, A., Grant, C., Demleitner, M., Henneken, E., & Murray, S. S. (2005). The effect of use and access on citations. Information Processing & Management, 41(6), 1395-1402, https://doi.org/10.1016/j.ipm.2005.03.010

Larivière, V., Sugimoto, C. R., Macaluso, B., Milojević, S., Cronin, B., & Thelwall, M. (2014). arXiv E-prints and the journal of record: An analysis of roles and relationships: arXiv E-Prints and the Journal of Record. Journal of the Association for Information Science and Technology, 65(6), 1157–1169. https://doi.org/10.1002/asi.23044

Maggio, L. A., Artino Jr, A. R., & Driessen, E. W. (2018). Preprints: Facilitating early discovery, access, and feedback. Perspectives on Medical Education, 7(5), 287–289. https://doi.org/10.1007/s40037-018-0451-8

Moed, H. F. (2007). The effect of "open access" on citation impact: An analysis of ArXiv's condensed matter section. Journal of the American Society for Information Science and Technology, 58(13), 2047–2054. https://doi.org/10.1002/asi.20663

Piwowar, H., Priem, J., Larivière, V., Alperin, J. P., Matthias, L., Norlander, B., … Haustein, S. (2018). The state of OA: a large-scale analysis of the prevalence and impact of Open Access articles. PeerJ, 6, e4375. https://doi.org/10.7717/peerj.4375

Powell, K. (2016). Does it take too long to publish research? Nature, 530(7589), 148–151. https://doi.org/10.1038/530148a

Ruocco, G., Daraio, C., Folli, V., & Leonetti, M. (2017). Bibliometric indicators: the origin of their log-normal distribution and why they are not a reliable proxy for an individual scholar's talent. Palgrave Communications, 3, 17064. https://doi.org/10.1057/palcomms.2017.64

Serghiou, S., & Ioannidis, J. P. A. (2018). Altmetric Scores, Citations, and Publication of Studies Posted as Preprints. JAMA, 319(4), 402. https://doi.org/10.1001/jama.2017.21168

van der Loo, M. P. J. (2014). The stringdist Package for Approximate String Matching. The R Journal, 6(1).

Wilcoxon, F. (1945). Individual comparisons by ranking methods. Biometrics Bulletin, 1 (6), 80-83. https://doi.org/10.2307/3001968

The Dynamics of French publications in Social Sciences and Humanities: A European comparison

Aouatif de La Laurencie[1] and Abdelghani Maddi[2]

[1]*aouatif.de-la-laurencie@hceres.fr*
Observatoire des Sciences et Techniques, Hcéres, Rue Albert Einstein, Paris, 75013 (France)

[2]*abdelghani.maddi@hceres.fr*
Observatoire des Sciences et Techniques, Hcéres, Rue Albert Einstein, Paris, 75013; CEPN, UMR-CNRS 723, Université Paris 13 (France)

Abstract:
Since 2000, the growth of French SSH publications in the WoS has been slower than that of comparable non-English-speaking European countries. Our study details this slower growth and discusses its causes. Two hypotheses are examined. First, France would be specialized in SSH disciplines that publish few articles. This could contribute to explain why French production is less visible in the international databases that mainly index articles. Second, French SSH scholars would publish relatively more in their national language than other European scholars, and would thus be less represented in journals in English. Our results indicate that in the SSH disciplines in which France is specialized, the proportion of articles is relatively low. Yet, France publishes a larger proportion of articles than the world average in these disciplines. The second hypothesis is confirmed to the extent that France has the lowest share of publications in English within the benchmark group of European countries. This remains true despite the fact that French publications in English are growing more strongly than those in French. The analysis also suggests other explanations that should be further explored.

Introduction

Quantitative measurements and the use of bibliometric indicators have increasingly become the standard in scientific research evaluation. However, many French scholars in social sciences and humanities (SSH) criticize these indicators, calculated from international citation databases such as Web of Science (WoS) or Scopus, claiming that these indicators underestimate their actual research production. This criticism stems from the coverage issues of these databases. First, they mainly list articles and proceedings at the expense of other document types (monographs, books and book chapters). Second, publications written in English are overrepresented in these bases.

Logically, coverage biases should penalize SSH in all non-English-speaking countries. Yet, Germany, Italy and Spain, which are comparable to France in terms of scientific research, have sharply increased their SSH publications in international databases over the last fifteen years. In contrast, French publications have grown slowly, although the country has a large scientific community in these disciplines.

Our paper aims to analyze the reasons for the slow growth of French SSH publications within WoS in comparison to this group of non-English-speaking European countries. To do so, we examine two hypotheses related to biases cited above.

Our first hypothesis is that France is specialized in SSH disciplines that publish few articles. This would explain why the scientific French production in these disciplines is less visible in the international databases that mainly list articles.

Our second hypothesis can be formulated as follows: French SSH scholars publish in French, rather than English, unlike scholars in the group of non-English-speaking European countries who tend to publish more in English. French publications are therefore less visible.

To verify these two hypotheses, we firstly identified SSH disciplines in which France is specialized. We then analyzed the distribution of document types by discipline for France in WoS and compared it to the World. Next we compared the French share of articles in WoS

and in the French open repository HAL. Secondly, we analyzed how the share of publications in English evolved from 2000 to 2015 in France, Germany, Italy and Spain. We added the Netherlands regardless of its volume of publications.
The first section of this paper presents a literature review on the specificities of SSH and their representation in international databases. The second section explains our methodology. The third section frames the analysis by presenting the progress in volumes and specializations of the selected countries. Sections four and five verify respectively our two hypotheses presented above. Finally, the discussion of the results raises new questions as tracks for future research.

Literature review: Specificities of SSH and Database Coverage

Specificities of the social sciences and humanities

The issue of the specificities of social sciences and humanities (SSH) is widely discussed in the literature. Hicks (1999, 2004) defines four types of research outputs in these disciplines. First, books play a central role in knowledge diffusion (unlike in sciences or engineering). However, they come in second place to articles published in peer-reviewed journals in terms of use in bibliometric analysis, because their coverage in international databases, like WoS or Scopus, is limited (Waltman 2016). Articles with a national orientation constitute the third type of output in SSH. They are generally published in national journals with smaller readerships. Their academic impact is relatively small, but their social impact can be high (Gruzd et al. 2013; Mohammadi et al. 2013, 2014, 2015a, 2015b; Mohammadi, 2014; Chen et al. 2015). The same is true of non-academic literature (intended for a wide public, such as documentaries, press articles, scientific popularization conferences and seminars), which is the fourth type of output in SSH defined by Hicks.
Hence, some SSH disciplines show similar behaviors to sciences. For instance, in economics, management, psychology and library & information science, articles play an important role (Hicks, 2004, Huang et al, 2008). According to McDonald (2003), the same is true of philosophy. In contrast, in history and literature, articles are relatively less used. In a sample of 173 publications totaling nearly 12,000 citations in three disciplines (philosophy, history and literature), McDonald (2003) shows that books account for more than 59% of all citations. There is a significant difference between the three disciplines: in philosophy, 44% of citations refer to books and 54% to articles, while in history and literature, these shares are respectively 57%/39% and 78%/16%.
Knievel and Kellsey (2005) provide a comparative analysis of eight disciplines: philosophy, music, religion, literature, linguistics, classics, philosophy and history. Their analysis covers more than 9,000 references to articles published in 2002. They show that within the human sciences there are divergences in citation practices. On average, book citations represent 76%. This rate varies between 51% (philosophy) and 89% (religion). There are disciplines in which articles published in journals play a central role in knowledge diffusion, including philosophy (51%) and linguistics (61%). Larivière et al. (2006) show that within SSH, the share of articles is increasing: it rose from 40 to 48% between 1981 and 2000, but more specifically from 60 to 70% in psychology and from 45 to 57% in economics. For other disciplines the rate is relatively stable. History, literature and the "other humanities" cite articles much less, with rates of 34%, 26% and 21% respectively.
Heinzkill (2007) analyzes 20,000 references from 555 articles published in literary journals. His results indicate that book citations represent more than 78%, of which 45% are more than 20 years old. In this case, it must be borne in mind that some of these citations refer to the original books studied, which is a peculiarity of literary studies.
In sum, SSH research outputs can be divided into four main groups: books, scientific articles conference proceedings, communications and seminars; and finally, science popularization products. Books are particularly important in some disciplines.

Coverage of databases: what place for SSH?

Database coverage level is the main condition of validity in bibliometric analyses. Well-known international databases like Web of Science (WoS) and Scopus mainly cover articles published in peer-reviewed journals and less the other scientific outputs like books (Nederhof, 2006 ; Mongeon et al. 2016). This can be problematic, especially for some social sciences and humanities disciplines in which articles are less widely used in knowledge diffusion.

According to Leydesdorff (2003), 55% of bibliographic references in social sciences are not indexed in the Social Sciences Citations Index (SSCI) of WoS. In the sciences, the percentage of coverage of bibliographic references in the Science Citation Index (SCI) is significantly higher, reaching almost 80%. It is worst for books. Butler's (1998) data on Australian social sciences indicate a strong and negative correlation (-0.83) between the proportion of publications indexed in the SSCI and the share of books in a discipline.

Mongeon and Paul-Hus (2016) compare WoS and Scopus coverage using the bibliographic database Ulrich as a reference on disciplinary, geographical and linguistic aspects. Out of a total of 63,013 academic journals indexed by Ulrich, WoS covers 13,605 journals, while Scopus lists 20,346. Both WoS and Scopus contain geographical and linguistic biases with an overrepresentation of English-language journals and countries. They also contain disciplinary biases, as both databases mostly include articles, which can be problematic, particularly in the humanities and social sciences80% of French journals in the Ulrich database in social sciences are not indexed by WoS. This rate is about 58% in arts and humanities, 43% in natural sciences and engineering, and 60% in biomedical research. The same applies to French-language journals: Scopus has a wider coverage of French-language publications than WoS.

Harzing et al. (2017a) and Harzing et al. (2017b) compared the coverage of four bibliometric databases (WoS, Scopus, Google Scholar and Microsoft Academic). Their sample consists of publications from 145 senior researchers representing the five main fields of science (Life Sciences, Sciences, Engineering, Social Sciences, and Humanities). They conclude that Microsoft Academic is an "excellent alternative" for bibliometric analyses. However, they point out that book coverage needs to be improved in Microsoft Academic, especially in SSH, as well as the quality of metadata.

Hug and Brändle (2017) provide a much broader comparative study by analyzing the coverage of three databases, Microsoft Academic, WoS and Scopus. They analyze the coverage of Zurich University researchers' publications in the open archive, "Zurich Open Archive and Repository" (ZORA), created in 2006. They analyze a subset of the ZORA database consisting of 62,791 publications published between 2008 and 2015. Their results show that the three databases have similar coverage rates. Yet, Microsoft Academic has better coverage of social sciences and humanities than the other two databases.

In summary, SSH are the least represented with relatively low coverage rates compared to those of the sciences. This is due in large part to the specificities of the social sciences and humanities, which have a range of scientific products that is not limited to articles published in peer-reviewed journals, and more national and regional audience.

Methodology

In this study, we use two data sources: WoS and HAL. The former is the well-known citation database commercialized by Calarivate Analytics. The data has been extracted from Observatoire des sciences et Techniques' (OST), in-house database. The database includes five indexes of the Web of Science (WoS) available from Clarivate Analytics (SCIE, SSCI, AHCI, CPCI-SSH and CPCI-S). As we started our study at 2017, the data corresponds to the full WoS content indexed until 2015. The second data source is HAL, which is the open

French repository launched in 2001 by CNRS (Centre national de la recherche scientifique). Scholars file their productions on a voluntary basis. We extract and analyze data about SSH deposits in HAL to compare them with data from WoS.

To understand why France's SSH publications in WoS have grown slowly, unlike those of comparable non-English speaking European countries, and given what the scientific literature tells us about the coverage issues of citation databases, we make two hypotheses.

Firstly, we hypothesize that France is specialized in SSH disciplines that publish few articles and so, since WoS lists mainly journal articles, French production is under-represented in this database. To verify this hypothesis, we first identify in which SSH disciplines France is specialized. Secondly, we analyse the distribution of document types for France and for the World. If France is specialized in SSH disciplines that publish mainly articles at the world level in WoS, France can be said to be underperforming. Conversely, if these disciplines publish few articles at the world level, France is disadvantaged by the overrepresentation of articles in WoS.

To identify France's disciplines of specialization, we use the OST disciplinary nomenclature that classifies WoS categories into 11 major disciplines. In this nomenclature, there are the major discipline "Humanities" and the major discipline "Social Sciences. We group the 81 SSH categories in WoS into 15 clusters as described in table 1.

Table 1: SSH clusters from WoS categories

Social Sciences	Humanities
Health	Arts
Information science and communication	History, archaeology
Sociology, demography, anthropology	Language, linguistics
Law, political science	Literature
Economics	Philosophy, ethics
Education	Psychology
Business, finance, management	-
Geography, urban studies, architecture	-
Multidisciplinary	

We then calculate France's specialization index for these clusters and compare it to the world average to find out in which clusters France is specialized.

Once we identify France's specialization clusters, we calculate the share of articles in World and French SSH publications in WoS and compare them. Lastly, we compare France's share of articles in WoS and in HAL.

Our second hypothesis takes into account the linguistic bias in WoS. It can be formulated as follows: French SSH scholars publish in French, rather than English, unlike scholars in the group of non-English-speaking European countries who tend to publish more in English. French publications are therefore less visible. To verify our second hypothesis, we first explore the evolution of publication volumes in English and in the national languages of the 5 countries in our group, namely France, Spain, Italy, Germany and the Netherlands. In a second step, we analyze the evolution of the share of English in SSH publications for each of these countries.

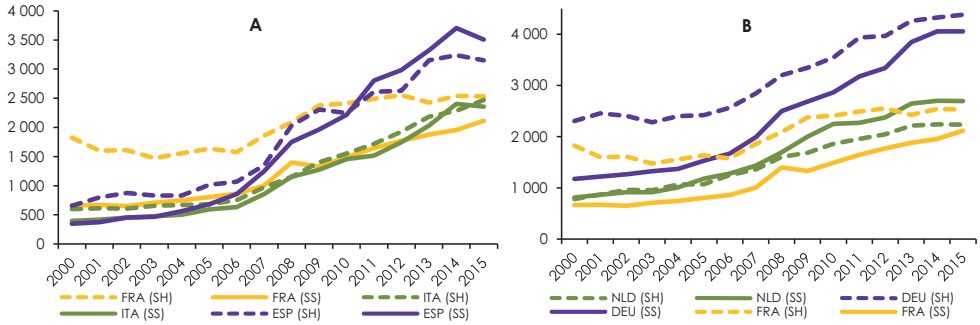
SSH Overview of the group of non-English-speaking European countries

The aim of this section is to show the difference between the evolutions of SSH publications in WoS as well as the disciplinary specialization of our benchmark group.

Number of SSH publications in WoS

At the world level, the number of publications referenced in WoS has almost doubled in Humanities (320,536 publications in 2015 against 180,994 in 2000). Publications in social sciences (SS) increased from 124,903 to 378,640 over the same period.

Figure 1A shows the impressive growth of Spanish SSH publications. In 2015, Spain published more than 3,500 papers in SS and more than 3,000 in Humanities. Italy multiplied its publications by 7 over the period 2000-2015. In contrast, France recorded slow growth and has been overtaken by Spain in both SS and Humanities and by Italy in SS since 2012.



Source : OST report, « La position scientifique de la France dans le monde 2000-2015 »

Figure 1: Number of SSH publications in WoS, 2000-15

Figure 1B shows that the gap between Germany and France in SS widened between 2000 and 2015. This is also the case for Humanities. Germany, with a relatively large volume of publications, grew more rapidly than France over the period. The same remark is true for the Netherlands, which has overtaken France since 2000 in social sciences. In contrast, the Netherlands had a lower volume of publications in Humanities than France.

Specialization index

We have seen in the section above that the number of SSH publications from France increased less than that of the other benchmark countries from 2000 to 2015. We will see that the same trend is true for the specialization index of the countries of interest.

The specialization index of a country in a discipline is the ratio between the share of the discipline in the country's publications and the share of the same discipline in world publications. The higher above 1 the specialization index is, the more the country is specialized in the discipline in question.

Table 2 shows that, in Humanities, very strong growth of the specialization index is observed for Spain (from 0.72 to 1.46). The Netherlands also strengthened its specialization index, as did Germany. France's specialization index remained stable over the period (0.95 in 2000 against 0.97 in 2015). Italy, despite having a very low specialization index in 2000, caught up with France in 2015.

In Social Sciences, Spain's specialization index moved from 0.40 to 1.17. France increased its specialization index in SS by 66% but remains below the world average. This is also the case for Germany. Italy recorded stronger growth, just behind Spain. The Netherlands, already specialized in 2000, reinforced its specialization in 2015.

Table 2: the specialization index in SS and Humanities in 2000 and 2015

	Specialization index in Humanities			Specialization index in SS		
Country	**2000**	**2015**	**Growth**	**2000**	**2015**	**Growth**
NLD	1,09	1,80	+66%	1,18	1,55	+32%
ESP	0,72	1,46	+104%	0,40	1,17	+194%
DEU	0,87	1,16	+32%	0,46	0,77	+66%
FRA	0,95	0,97	+2%	0,36	0,58	+61%
ITA	0,47	0,96	+105%	0,32	0,65	+103%

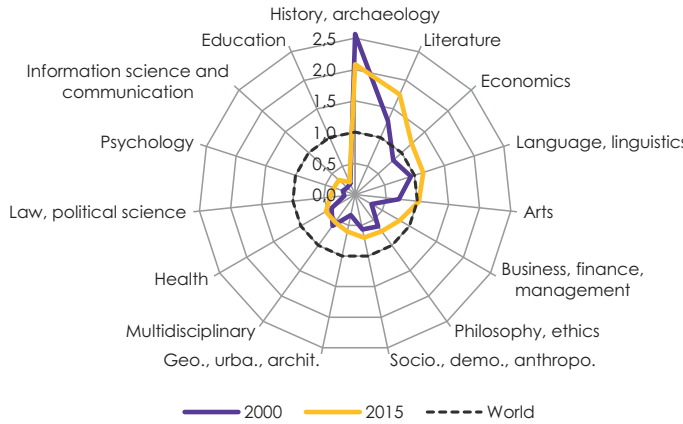
Source: OST base, Web of Science, OST calculations

The biases in coverage of international databases such as WoS seem to be more critical for France than the other countries France shows limited growth of specialization index. Therefore, in the following two sections, we focus only on the French case and verify our two hypotheses: about document types that SSH publish and about publication languages.

Specialization profile and France's SSH publications

Document types in WoS

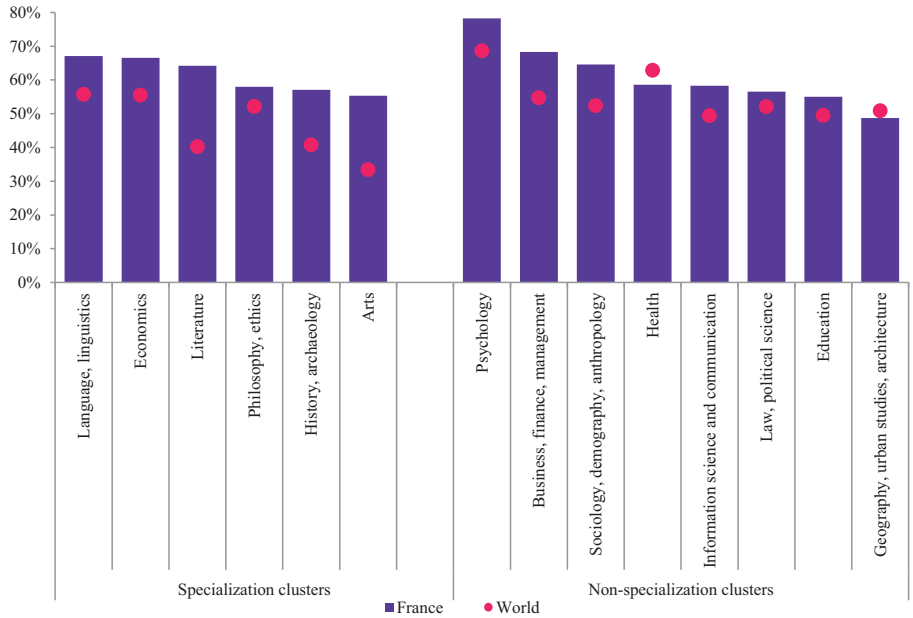
To explain its slow growth in WoS, we hypothesize that France is specialized in SSH disciplines that publish few articles and, since WoS lists mainly journal articles, French production is under-represented in this database. To verify this hypothesis, we first need to identify in which SSH disciplines France is specialized and then characterize French publications in terms of document types. To do so, we calculate France's specialization index for the clusters in SSH (see table 1) and compared it to the world average to find out in which clusters France is specialized. Figure 2 provides data about France's specialization index in 2000 and 2015 and shows that France is specialized more in Humanities than in Social sciences disciplines. Within the former discipline, France was specialized in history-archeology (2.08), in literature (1.75), in language-linguistics (1.14) and in arts (1.02) in 2015. In the social sciences, France became specialized in economics (index of 1.2 in 2015).



Source: OST base, Web of Science, OST calculations

Figure 2: Specialization index of France in SSH clusters. 2000-15

Figure 3 shows the share of articles in the SSH specialization and non-specialization of France in 2015.



Source: OST base, Web of Science, OST calculations

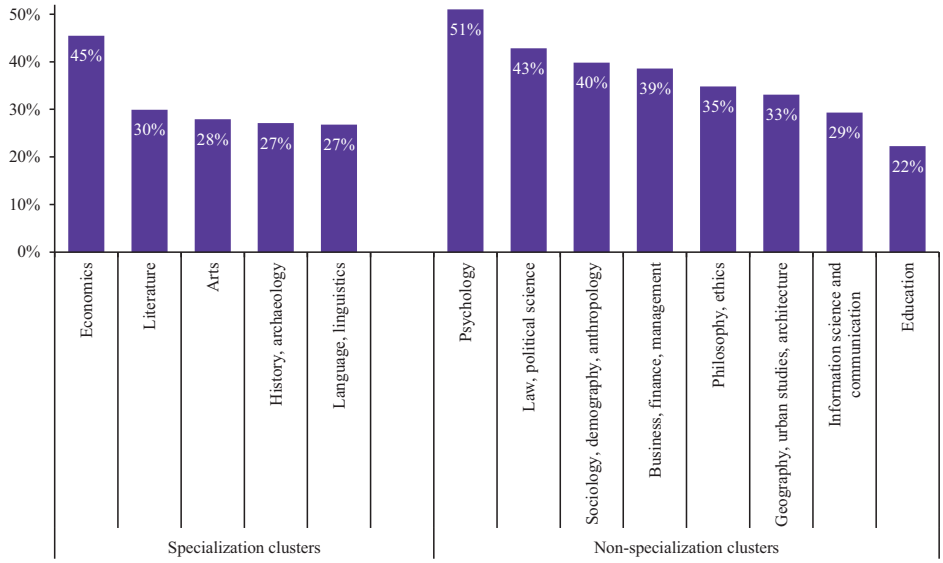
Figure 3: Share of articles in specialization and non-specialization SSH clusters of France in 2015 in WoS

At the global level, the share of articles in language-linguistics cluster and in economics is 56% in WoS. Conversely, the world share of articles is small in arts (33%), in letters (40%) and in history-archeology (41%). The share of articles among French publications in WoS is much higher in all specialization clusters. 67% of French publications in languages-linguistics and economics are articles, 64% in letters, followed by history-archeology and arts respectively with 57% and 55%. A similar trend is observed in France's non-specialization clusters of France, except in the health and geography, urban studies and architecture clusters.

Comparison between WoS and HAL

As a second step, we compared the distribution of document types in French specialization and non-specialization clusters in WoS and in HAL, the French open repository to check whether France's clusters of specialization and non-specialization behave in the same way in WoS and HAL.

Figure 4 shows the share of articles in HAL for France's specialization and non-specialization clusters in 2015. In France's specialization clusters, the share of articles does not exceed 30% (except in economics). Psychology is the discipline with the highest share of articles in HAL (51%).



Source : HAL, OST calculations

Figure 4: share of articles in France's specialization and non-specialization SSH clusters in 2015 in HAL

The analysis of WoS and HAL data makes it possible to partially verify our first hypothesis: The clusters where France is specialized are not dominated by articles. Paradoxically, France publishes more articles than the world average in all these clusters. These results could be explained in two ways. Firstly, other French publications (books, book chapters, etc.) are mostly published in French without any abstract in English or without a translation in English. Therefore they not listed in WoS. Secondly, France's scientific production in SSH is simply small in articles intensive disciplines as Psychology compared to that of other countries, and France is not visible in WoS.

Language of publications

In this section, we first explore the evolution of publication volumes in English and in the national languages of France, Spain, Italy, Germany and Netherlands. In a second step, we analyze the evolution of the share of English in the SSH publications of each of these countries.

The aim is to verify whether the under-representation of French SSH in WoS can be explained by the fact that the French scholars publish more in their national language, whereas international bibliometric databases tend to cover publications in English.

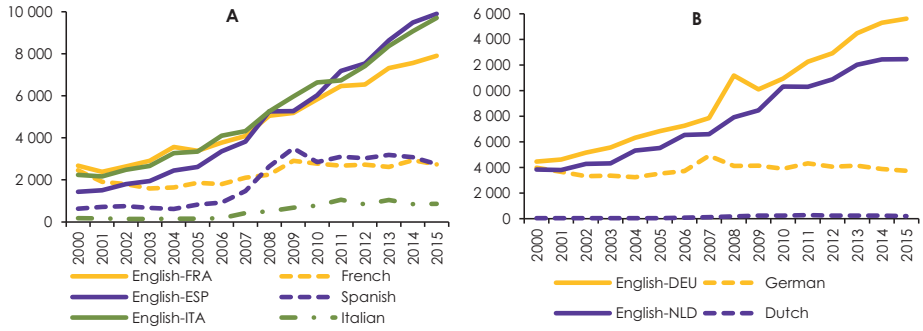


Figure 5: Publication numbers in SHS by language

Figure 5A shows the evolution of publications in English and in national languages. In 2000, the number of publications by France exceeded that of the other two countries. Italy and Spain have since recorded stronger growth than France, and exceeded it from 2005 (for Italy) and 2008 (for Spain). Spain's growth is so strong that it also overtook Italy in 2010. In contrast, Italy publishes less in its national language than France and Spain. Like figure 5A, figure 5B shows a big increase in English-language publications for both Germany and the Netherlands. In contrast, national language publications continue to decline for Germany, and are marginal for the Netherlands.

As for national languages, figure 6 shows an interesting trend. Spain strongly increased its volume of publications in its national language, Spain's publications count for almost 50% of world publications in Spanish, especially between 2006 and 2009. This trend decelerated from 2010, but the volume of Spain's publications in the national language and slightly overtook the France's publications in French that increased slowly during this period. It is worth noting that France's publications count for 65% of world publications in French the same slow trend can be observed for Germany while national language publications are marginal for the Netherlands and Italy.

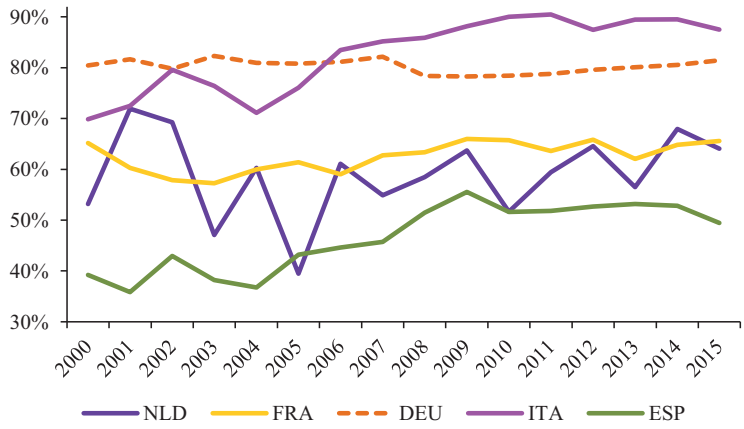


Figure 6: countries' contribution in publication in SHS by national language

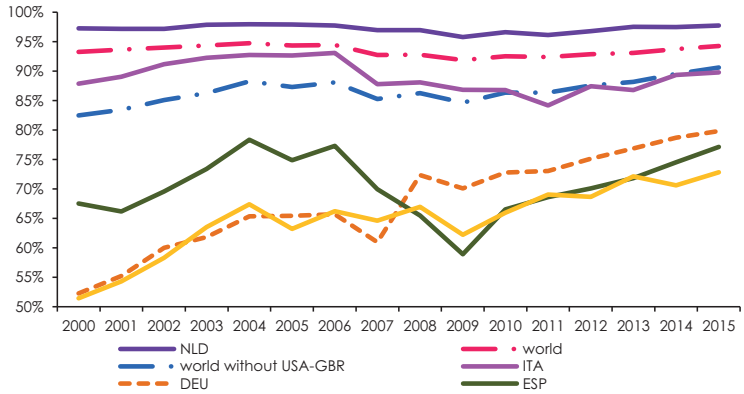


Figure 7: Share of English in the national SSH publications of the group of countries

Figure 7 shows the evolution of the share of English language SSH publications of our group. For the Netherlands and Italy, the share is very high (90-100% over the period). An exceptional trend is also observed in the case of Germany, which in 2000 had just over 50% of its publications in English; the share reached 80% in 2015. The case of Spain is particular; its share of English-language SSH publications increased from 65 to 75% between 2000 and 2006. Then the share dropped until 2009, when it dipped below 60%. This period corresponds to the strong growth in the volume of Spanish-language publications (see figure 5A). After 2009, the share of English in Spanish publications rose again and now exceeds that of France. France has also increased its share of publications in English. However, given its trend, it is the country with the lowest share (73%) within our benchmark group.
With regard to the two previous figures, our hypothesis on the language of publication is verified. France is the country with the lowest share of publications in English. And conversely, it has the highest share of publications in the national language. However, the strong growth of Spain's national language publications raises questions. In particular, why are Spanish publications more visible in WOS than publications in French? Is this an artifact of the database that has increasingly integrated Spanish journals?

In fact, the number of journals from Spain in WoS rose from 31 to 105 between 2000 and 2015. We make the hypothesis that the rose of journals number is correlated with the rise of Spain's publications. We also assume that the Spain visibility in WoS profited from the increase of the number of journals from Latin and South America integrated in the database (from 23 to 105 during the same period). In contrast, the number of journals from France stagnated in WoS around 122 journals with only +5% of growth between 2000 and 2015. Again, we make the hypothesis that this stagnation has a negative impact on France visibility in WoS. In fact, it would be interesting to analyze this question in depth because Spanish or French scholars don't necessarily publish in national journals and could decide to publish in other journals regarding their international visibility or thematic specialisation.

Conclusions and discussion

The purpose of this work is to analyze the dynamics of French social sciences and humanities publications in the WoS database by comparing them with those of the comparable non-English-speaking European countries. We started from the hypothesis that France may be doubly penalized by the overrepresentation of articles and English in international bases.

The evolution of SSH publications in the WoS database is highly contrasting in the benchmark group. France is the country to have made the least progress over the period 2000-2015. Two hypotheses are likely to explain this low progression by France. The first is that France is specialized in some SSH where the share of articles is relatively low. As international databases cover mostly articles, these disciplines are not well represented.

The results indicate that in specialization clusters of France, the proportion of articles is relatively low except in linguistics and economics. This verifies partially our first hypothesis as France publishes more articles than the world average in these clusters. This result raises several questions: do only French articles get through the barrier at the entrance to WoS? Is it easier for French scholars to publish an article in English than a book, given that the time required for preparing and writing a SSH book is longer? Is French production small in articles intensive SSH disciplines?

Regarding the second hypothesis, French scholars in SSH publish less in English than those in other non-English-speaking countries. French publications in English are growing more strongly than French language publications. Nevertheless, other countries have recorded a much larger increase in publications in English. Spain is a particular case as it has increased also its publications in its national language in WoS. France is the last country in terms of share of publications in English; it has difficulty asserting its presence in WOS.

Thus, language may partially explain the low visibility of French SSH in WOS, but questions remain unanswered, especially regarding Spain. Why did Spain have such a strong dynamic of publications written in Spanish in WoS? Is it an artifact of the database that favors the Spanish language (widely used in Latin America) or has it adopted a catch-up strategy for Spain? Or is it more a genuine progression by Spain as a corollary of an increase of R&D expenditures or an efficient research strategy making scholars more productive? The OST report (2018) showed that Spain is the first country to have benefited from the integration (or creation) into WoS of new journals in the social sciences for instance. It would be interesting to analyze this question in depth in order to find out what has afforded Spain such a progression.

In summary, it is partially possible to answer our question about the low presence of French SSH in WOS. It does not seem to be due solely to poor SSH coverage by the database as there are significant differences between non-English-speaking countries. The lack of adaptation to international standards of publication, notably publishing articles in English or at least including an abstract in English, may constitute a part of the explanation. It would also be worth conducting, in the future, an in-depth analysis of national characteristics of SSH publications and of public policies related to scientific production and publication practices in non-English-speaking countries as Spain , Italy and also Belgium and their impact on the visibility.

References

Butler, L. (1998). Personal communication of unpublished data: Cf. Hicks (2004), 479.

Chen, K., Tang, M., Wang, C., & Hsiang, J. (2015). Exploring alternative metrics of scholarly performance in the social sciences and humanities in Taiwan. Scientometrics, 102(1), 97–112. http://doi.org/10.1007/s11192-014-1420-6.

Glänzel W., Schoepflin, U. (1999). A bibliometric study of reference literature in the sciences and social sciences. Information Processing and Management, 35, 31-44.

Gruzd, A., & Goertzen, M. (2013). Wired Academia: Why Social Science Scholars Are Using Social Media (pp. 3332–3341). Presented at the 46th Hawaii International Conference on System Sciences, Maui, Hawaii: IEEE. http://doi.org/10.1109/HICSS.2013.614.

Harzing, A. W., & Alakangas, S. (2017a). Microsoft Academic: Is the phoenix getting wings? Scientometrics, 110(1), 371–383.

Harzing, A. W., & Alakangas, S. (2017b). Microsoft Academic is 1 year old: The Phoenix is ready to leave the nest. Scientometrics, 112(3), 1887–1894.

Heinzkill, R. (2003). References in scholarly English and American literary journals thirty years later: A citation study. COLLEGE & RESEARCH LIBRARIES. Volume: 68. Issue: 2. Pages: 141-153.

Hicks, D. (2004). The four literatures of social science. In H. F. Moed, W. Glänzel, & U. Schmoch (Eds.), Handbook of quantitative science and technology research: The use of publication and patent statistics in studies of S&T systems (pp. 476–496). Dordrecht: Kluwer Academic Publishers.

Hug, S. E., & Brändle, M. P. (2017). The coverage of Microsoft Academic: Analyzing the publication output of a university. Scientometrics. https://doi.org/10.1007/s11192-017-2535-3.

Katina Strauch (Westport, Conn.: Libraries Unlimited, 2003), 144.

Knievel, J.E., & Kelley, C. (2005). Citation analysis for collection development: A comparative study of eight humanities disciplines. The Library Quarterly 75(2): 142-168.

Larivière et al (2006). The place of serials in referencing practices: Comparing natural sciences and engineering with social sciences and humanities. JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE AND TECHNOLOGY. Volume: 57. Issue: 8. Pages: 997-1004.

Leydesdorff, L. (2003). Can networks of journal–journal citation be used as indicators of change in the social sciences? Journal of Documentation, 59 (1), 84–104.

McDonald, J. (2003). The Book is Dead," in Charleston Conference Proceedings, 2001, ed.

Mohammadi, E. (2014). Identifying the Invisible Impact of Scholarly Publications: A Multi-disciplinary Analysis Using Altmetrics (Ph.D. thesis). University of Wolverhampton, Wolverhampton, UK.

Mohammadi, E., & Thelwall, M. (2013). Assessing non-standard article impact using F1000 labels. Scientometrics, 97(2), 383–395. http://doi.org/10.1007/s11192-013-0993-9.

Mohammadi, E., & Thelwall, M. (2014). Mendeley readership altmetrics for the social sciences and humanities: Research evaluation and knowledge flows. Journal of the Association for Information Science and Technology, 65(8), 1627–1638. http://doi.org/10.1002/asi.23071.

Mohammadi, E., Thelwall, M., Haustein, S., & Larivière, V. (2015a). Who reads research articles? An altmetrics analysis of Mendeley user categories. Journal of the Association for Information Science and Technology. http://doi.org/10.1002/asi.23286.

Mohammadi, E., Thelwall, M., & Kousha, K. (2015b). Can Mendeley Bookmarks Reflect Readership? A Survey of User Motivations. Journal of the Association for Information Science and Technology, http://doi.org/10.1002/asi.23477.

Mongeon, P., & Paul-Hus, A. (2016), The journal coverage of Web of Science and Scopus: A comparative analysis. Scientometrics, 106(1), 213-228.

OST (2018), La position scientifique de la France dans le monde, 2000-2015, Hcéres, Paris.

Sugimoto, C. R., Work, S., Larivière, V., & Haustein, S. (2017). Scholarly use of social media and altmetrics: A review of the literature. *Journal of the Association for Information Science and Technology*, 68(9), 2037–2062.

Waltman L. (2016), A review of the literature on citation impact indicators. *Journal of infometrics*. 10, 365–391 (2016). doi: 10.1016/j.joi.2016.02.007.

The spatial distribution of knowledge production in Europe. Evidence from KET and SGC

Benedetto Lepori[1], Massimiliano Guerini[2], Thomas Scherngell[3] and Philippe Laredo[4]

[1] blepori@usi.ch
Università della Svizzera Italiana, Via Lambertenghi 10A, 6900 Lugano (Switzerland)

[2] massimiliano.guerini@polimi.it
DIG - Politecnico di Milano, Via Lambruschini 4b, 20156 Milano (Italy)

[3] thomas.scherngell@ait.at
Austrian Institute of Technology, Giefinggasse 4, 1210 Vienna (Austria)

[4] philippe.laredo@enpc.fr
Université Paris-Est Marne-la-Vallée, 5 Boulevard Descartes, 77454 Champs-sur-Marne (France)

Abstract

In this paper, we develop an analysis of the spatial distribution of knowledge production related to Key Emerging Technologies (KETs) and Societal Grand Challenges (SGCs) in Europe building on an extensive dataset developed in the H2020 KNOWMAK project. We first provide a broad characterization of European regions in terms of their knowledge volume and knowledge intensity, which leads to a distinction between the large metropolitan regions and smaller knowledge intensive regions. Second, by using principal component analysis, we identify two components of knowledge production that we broadly characterize as academic production and technology production. This distinction allows further categorizing regions in terms of the balance between the two components, which we suggest is also related to the ecology of actors in a region and, notably, of the importance of public-sector research and of knowledge producing firms. In a further step, we will adopt more advanced statistical techniques, i.e. latent class analysis, in order to provide a robust identification of classes of regions.

Introduction

The creation of new knowledge is the essential basis for successfully generating innovation, and thus, described as a major determinant of the overall socio-economic development of regions or countries (Audretsch and Feldman 1994). In this context, the investigation of the distribution and diffusion of knowledge in geographical space, and how this distribution changes over time, has become one of the main research domains in Regional Science and Economic Geography (Feldman and Kogler 2010), and also gaining increasing interest in Science, Technology and Innovation (STI) studies. Most empirical works in this direction employ a regional perspective, pointing to the importance of different regional-internal and regional-external characteristics for a region´s ability to create new knowledge, and, by this, to gain competitive advantages (see, e.g., Scherngell 2013, Wanzenböck et al. 2014).

However, looking at previous literature, we find that in empirical terms most studies concentrate on the characterization of the wider regional ecosystem for innovation (see e.g. Moreno et al. 2006, Navarro et al. 2009, Verspagen 2010), mostly focusing on innovation related indicators, while the underlying knowledge base is often underemphasized, in particular the scientific knowledge base. Usually, indicators on patenting are exclusively used to describe a regions knowledge creation capability, often meshed with other innovation related indicators (e.g. human ressources) in a linear-additive manner to a synthetic index (see e.g. Hollanders et al. 2009, Dunnewijk et al. 2008). Moreover, these works – most of them done for the European territory – employ regional breakdowns (such as NUTS2 for the European case) that often intersect agglomerations of knowledge creation, leading to problematic interpretations in a spatial context. In this sense, there is a clear need for advancing this research domain by providing a richer and more in-depth empirical basis on different types of regional knowledge

bases, and to come to a more meaningful regional classification system to analyse the spatial distribution of new knowledge.
In this paper, we aim to address this research gap by advancing our understanding of the spatial distribution of knowledge production in Europe, leveraging on a rich dataset developed by the H2020 project KNOWMAK (knowmak.eu). Our main goal is therefore to analyze the spatial distribution of knowledge production across European regions by focusing on three dimensions: (i) the volume of knowledge production (absolute and relative to the population); (ii) the balance between a more 'academic' and a 'technological' component; (iii) the relative specialization of regions for what concerns key technological domains and emergent societal challenges. By intersecting these three dimensions, we also aim at developing a robust and multidimensional classification of European regions in terms of knowledge production.
The paper provides three major advances to the current literature. First, we are able to interlink a rich set of data on the different facets of knowledge production, including the more academic outputs (scientific publications), project-based collaborative (FP projects), and technological knowledge production (patents). Second, we introduce a regional definition that, while still consistent with EUROSTAT definitions, takes into account the geography of knowledge production. Third, to come up with a robust classifications of regions, we adopt advanced statistical methods, such as Latent Class Analysis (LCA).

Methods

Data

The data are derived from a rich dataset on knowledge production in Europe developed within the H2020 project KNOWMAK[1]. The dataset includes data on scientific publications derived from the Web of Science version at the University of Leiden (Waltman, Calero-Medina, Kosten, et al 2012), on European collaborative R&D projects from the EUPRO database developed at the Austrian Institute of Technology (Roediger-Schluga and Barber 2008), and on patents from the PATSTAT version at IFRIS in Paris (Laurens, Le Bas, Schoen, Villard and Larédo 2015).
The perimeter of knowledge production corresponds to Key Enabling Technologies (KET[2]) and Societal Grand Challenges (SGC[3]) as defined by the European Union. From a policy perspective, KET can be considered as the emergent frontier of knowledge production, and SGC as the knowledge domains specifically crucial for the major societal challenges of the future. Data have been attributed to KETs and SGCs through advanced text annotation relying on a newly developed ontology of science and technology (Maynard and Lepori 2017). In a further step, these data will therefore allow introducing topical specialisation as a further dimension to characterize the spatial distribution of knowledge production.
All source data have been geolocalised based on the authors' (publications), participants' (projects) and inventors' addresses (patents). This allows for a flexible attribution to regions. A new regional classification has been developed to address some issues of EUROSTAT NUTS regions[4]. More precisely, the classification includes EUROSTAT metropolitan regions (based on the aggregation of NUTS3-level regions) and NUTS2 regions for the remaining areas; further, a few additional centers for knowledge production, like Oxford and Leuven, have been singled out at NUTS3 level. The resulting classification is therefore more fine-grained than NUTS2 in the areas with sizeable knowledge production, but at the same time recognizes the central role of metropolitan areas in knowledge production. Since it is fully based on the aggregation of NUTS3 regions, regional statistics by EUROSTAT can still be used.

[1] http://knowmak.eu
[2] https://ec.europa.eu/programmes/horizon2020/en/area/key-enabling-technologies
[3] https://ec.europa.eu/programmes/horizon2020/en/h2020-section/societal-challenges
[4] https://ec.europa.eu/eurostat/web/nuts/background.

The geographical perimeter considered includes EU-28 countries, EA-EFTA countries (Iceland, Liechtenstein, Norway and Switzerland) and candidate countries (Albania, Former Yugoslav Republic of Macedonia, Montenegro, Serbia and Turkey) for a total of 553 regions. Data refer to year 2013.

Indicators

Analyses of knowledge production classically combine publications (as shared outputs of scientific activity) and patents (as published proprietary knowledge anticipating for commercial applications). We add to this information on 'knowledge in the making' by using on-going projects (funded by the EU-FP). We use three indicators for each: simple production counts (publications, priority patents, participations in FP projects), indicators of collaborative activities (linking metropolitan areas to the world: international co-publications, transnational patents) and indicators of potential value (top 10% cited papers, top 10% patent families, co-ordinations of FP projects).

We use two indicators of regional size: population and Gross Domestic Product in Purchasing Power Parities, both produced by EUROSTAT. Further indicators that will be introduced in future work include topical specialization indicators (based on KET/SGC), regional network centrality indicators and indicators characterizing the actors in the region.

Analysis

As a first step, we analyze descriptively the extent of knowledge production by region. To this aim, we build two composite indicators:

- The *knowledge production share* as the average of the regional share of publications, projects and patents.
- The *knowledge production intensity* as the knowledge production share divided by the regional share of population in the whole perimeter.

As a second step, we use Principal Component Analysis (PCA) in order to identify relationships between indicators and to single out the main dimensions differentiating regions. These results will also allow to identify dimensions and indicators for a more advanced classification. To this aim, we rely on advanced statistical methods, i.e. Latent Class Analysis (LCA; Muthén 2004). This class of models fits the distribution of a set of observed variables conditional to the observations belonging to non-observed (latent) classes. Compared with conventional clustering methods, latent-class clustering presents the advantage of being model-based (hence it can incorporate prior assumptions on classes and statistical distributions) and has been shown to provide much better results (Magidson and Vermunt, 2002).

Preliminary results

Our data show that regions with higher knowledge production volumes are mostly concentrated in large metropolitan regions, with Paris (in France), London (in the UK) and Munich (in Germany) that rank in the first three positions. The distribution of the volume of knowledge production appears highly skewed, with the first 10 regions (mostly large metropolitan regions with a population higher than 2M inhabitants) that account for more than 20% together. However, among these regions, only Munich ranks in the top 10 regions as to intensity. Paris and London, which are by far the most important regions in terms of volume, rank #64 and #167, respectively, in terms of knowledge production intensity. Their position is therefore largely accounted for by their sheer demographic and economic size.

On the other hand, medium-size metropolitan areas like Eindhoven (in the Netherlands), Vlaams-Brabant (Leuven – Belgium) and Uppsala (in Sweden) rank in the first three positions in terms of production intensity, while still having rather large volumes of knowledge production (ranking #10, #30 and #50 and respectively). This emphasizes the important role of

such medium-size regions that is likely to emerge even more clearly when analyzing specific research domains.
Non-metropolitan areas typically exhibit lower levels of knowledge production, except when they include university cities, like East Anglia (Cambridge), Berkshire, Buckinghamshire and Oxfordshire (Oxford), and Zuid-Holland (Leiden and Delft). These regions rank #19, #22, and #24, respectively, for knowledge production. Such areas also emerge when looking at knowledge production intensity, where they rank #19, #27 and #25, respectively.
We notice that Eastern European countries are generally characterized by low volumes of knowledge production, with the exception of large capital cities like Prague (in the Czech Republic), Warsaw (in Poland), and Budapest (in Hungary). Production intensity of regions in Eastern Europe is however generally lower. Ljubljana (in Slovenia) is a notable exception, ranking #34 in the overall level of knowledge intensity.
The PCA identifies two main components. Table 1 presents the factor loadings associated with these two principal components (after varimax rotation). Extracted components explain 94% of the total variance. Loadings whose absolute value is greater than 0.4 are in bold. Based on the loadings, the two components can be labeled as *academic production* and *technology production* respectively.

Table 1. Basic PCA results with factor loadings

Variable	Comp.1 Academic production	Comp.2 Technology production	Unexplained variance
N. of publications	**0.47**	-0.04	0.04
N. of international publications	**0.46**	-0.04	0.05
N. of publications in the top 10%	**0.46**	-0.05	0.06
N. of participations to EU-FP projects	**0.41**	0.08	0.09
N. of coordinations to EU-FP projects	**0.42**	0.05	0.11
N. of priority patent applications	0.07	**0.65**	0.04
N. of transnational priority patent applications	-0.05	**0.75**	0.04

Academic production mostly relates to the three measures that capture the production of scientific publications. Participation to EU-FP projects has been assigned to this first component as well. This result is hardly surprising as research active universities and research organizations have better chances to get funding from EU-FP projects. On the contrary, *Technology production* relates to the two measures that capture regional knowledge codified in patents.
Figure 2 shows regions according to their levels of academic and technology production. We report the names of the regions only if the values of both regional academic and technology production are higher than the 90th percentile of their distributions. Large metropolitan areas such as Paris, London Munich, Berlin and Barcelona exhibit high levels of both components, while London shows a greater propensity towards academic production. A similar pattern is associated to the university areas of Oxford and Cambridge. Furthermore, we also observe smaller urban areas such as Heidelberg and Lyon with a non-negligible level of both academic and technology production.
The two details show regions that are characterized by high levels of technology production, but moderate to low levels of academic production (figure 3a), respectively by high levels of academic production, but moderate to low levels of technology production (figure 3b). Eindhoven and Stuttgart are leading centers of technology production, while showing a moderate level of academic production. We also observe regions with quite high levels of

technology production, but low levels of academic production, such as the Mannheim-Ludwigshafen region, Grenoble and Regensburg.
The national capital regions of Rome, Prague, Lisbon and Athens are all characterized by low levels of technology production, while being leading centers of academic production. These regions feature a number of large universities are located in those regions, while their industrial structures are typically not technology oriented. Similar patterns are observed for smaller urban areas that are characterized by the presence of important universities, such as Vlaams-Brabant (Leuven), Bologna, and Gent.

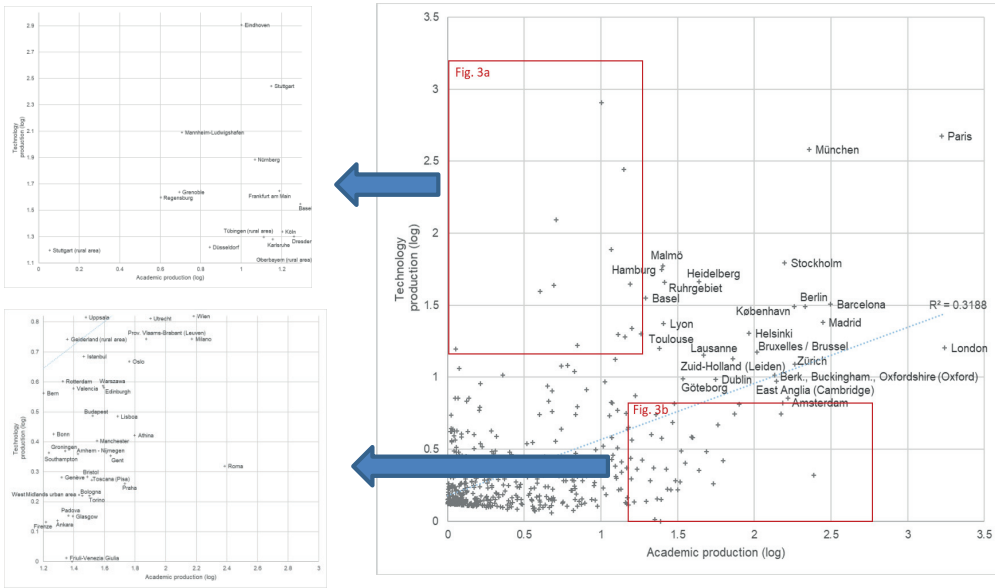


Figure 2. Scatter plot on academic and technology production (logarithmic scale)

Discussion and further work

This study contributes to the research stream investigating the geographical dimension of knowledge creation. Analysing the spatial distribution of different types of knowledge creation across a set of 553 metropolitan and other European regions, and relying on a rich set of interlinked data, we provide a novel picture on the European knowledge creation landscape.
Our preliminary results highlight two aspects. First, volume and intensity tend to differentiate strongly regions, as large production regions tend to have lower intensity than medium-sized regions. Relying on data by topic, we will test the hypothesis that this pattern is related to specialisation of topics vs breadth or ubiquity, i.e. the more specialised, the more productive.
Second, the PCA reveals the presence of different patterns of knowledge production across European regions and allows identifying a significant number of specialized regions in academic vs. technological knowledge production. By relying on the information on research actors, we will analyze how ‘anchor tenant’ actors play. More specifically we hypothesize that regions with one anchor tenant (a large firm or a large university) will exhibit unbalanced profiles, while regions with multiple anchor tenant actors will exhibit balanced profiles.
Clearly these preliminary results pave the way for future research endeavors. *First*, we will employ more advanced classification approaches that are inherently multidimensional and allow including complementary indicators. Latent Class Analysis (LCA) may be a promising

instrument in this context, allowing us to position regions against each other in terms of their knowledge creation characteristics. *Second*, making use of the rich underlying topical information on regional knowledge creation gives the possibility for novel specialization vs. diversification insights into regional knowledge bases. For instance, the breadth of a regional knowledge base is of crucial interest, not only in a research context, but also in a policy context, in particular in terms of the smart specialization debate. *Third*, an explanatory framework that tells us which region-internal and region-external determinants drive the observed spatial patterns of knowledge creation is high on the research agenda. In the latter context, a spatial econometric approach is most promising.

References

Audretsch, David B., and Maryann P. Feldman. "R&D spillovers and the geography of innovation and production." The American economic review 86.3 (1996): 630-640.

Dunnewijk, T., Hollanders, H. & Wintjes, R. (2008). Benchmarking regions in the enlarged Europe: diversity in knowledge potential and policy options. In Nauwelaers, C. & Wintjes, R. (Eds.), *Innovation Policy in Europe: Measurement and Strategy*. Edward Elgar, Cheltenham.

Feldman, M. P., & Kogler, D. F. (2010). Stylized facts in the geography of innovation. In R. Hall, N. Rosenberg (Eds.), *Handbook of The Economics of Innovation*, Elsevier, Oxford.

Hollanders, H., Tarantola, S. & Loschky, A. (2009). Regional innovation scoreboard (RIS) 2009. *Pro Inno Europe*.

Laurens, P., Le Bas, C., Schoen, A., Villard, L. & Larédo, P. (2015). The rate and motives of the internationalisation of large firm R&D (1994–2005): Towards a turning point? *Research Policy, 44(3)*, 765-776.

Marsan, G. A. & Maguire, K. (2011). Categorisation of OECD regions using innovation-related variables.

Maynard, D. G. & Lepori (2017). Ontologies as bridges between data sources and user queries: the KNOWMAK project experience.

Moreno, R., Paci, R. & Usai, S. (2006). Innovation clusters in the European regions. *European Planning Studies, 14(9)*, 1235-1263.

Muthén, B. (2004). Latent variable analysis. *The Sage Handbook of Quantitative Methodology for the Social Sciences,* , 345-368.

Navarro, M., Gibaja, J. J., Bilbao-Osorio, B. & Aguado, R. (2009). Patterns of innovation in EU-25 regions: a typology and policy recommendations. *Environment and Planning C: Government and Policy, 27(5)*, 815-840.

OECD (2008). Handbook on Constructing Composite Indicators. Methodology and user guide.

Roediger-Schluga, T. & Barber, M. (2008). R&D collaboration networks in the European Framework Programmes: data processing, network construction and selected results. *International Journal of Foreseight and Innovation Policy, 4(3-4)*, 321-347.

Scherngell, Thomas. "The Networked Nature of R&D in a Spatial Context." The Geography of Networks and R&D Collaborations. Springer, Cham, 2013. 3-11.

Verspagen, B. (2010). The spatial hierarchy of technological change and economic development in Europe. *The Annals of Regional Science, 45(1)*, 109-132.

Waltman, L., Calero-Medina, C., Kosten, J., Noyons, E., Tijssen, R. J., Eck, N. J., Leeuwen, T. N., Raan, A. F., Visser, M. S. & Wouters, P. (2012). The Leiden Ranking 2011/2012: Data collection, indicators, and interpretation. *Journal of the American Society for Information Science and Technology, 63(12)*, 2419-2432.

Wanzenböck, Iris, Thomas Scherngell, and Thomas Brenner. "Embeddedness of regions in European knowledge networks: a comparative analysis of inter-regional R&D collaborations, co-patents and co-publications." The Annals of Regional Science 53.2 (2014): 337-368.

Hedonic Pricing and the Valuation of Open Access Journals

Kyle Siler[1] and Koen Frenken[2]

[1] *ksiler@gmail.com*
University of Sussex - SPRU, Jubilee Building, BN1 9RH Falmer, Brighton (United Kingdom)

[2] *k.frenken@uu.nl*
Copernicus Institute of Sustainable Development, Princetonlaan 8a, 3584 CB Utrecht (Netherlands)

Abstract

The emergence of Open Access (OA) publishing has created new economic niches and debates in academic publishing. OA journals offer numerous publication outlets with differing editorial philosophies and business models. Scholars and academic stakeholders must decide which journals offer acceptable value for the direct or indirect costs of academic publishing. Our research uses the Directory of Open Access Journals (N=12,100) to identify various characteristics of OA academic journals that influence journal prices. The Journal Impact Factor (JIF), language, publisher mission, DOAJ Seal, World Bank Economic and Geographic regions of publishers, peer review duration and journal discipline are identified as factors with significant influence over journal price levels. Journals with status endowments (JIF, DOAJ Seal), published in wealthier regions, in medical or science-based disciplines, and with English-based articles are relatively more expensive. Scholarly and political economic inequalities manifest in the prices and benefits offered by different journals and publishers throughout the world.

Introduction

Online scholarly publishing has yielded numerous diverse economic and academic niches for Open Access (OA) journals. These new incentives and institutions shape pricing strategies for publishers, while influencing publication choices for scholars and academic stakeholders. OA scholarly publishing reduces the moral hazard with the subscription journal business model, where librarians and scholarly administrators tend to pay for journal subscriptions, instead of the primary consumers of the product (scholars). With OA journals funded via APCs (Article Processing Charges), prices are strategically set for individual journals, in contrast to subscription-based journals that are often paid for via "big deal" consolidations of large journal bundles. These factors contribute to the OA publishing market being relatively competitive and price-sensitive, since researchers often pay APCs to publish their research out of their own limited professional – and in some cases, personal – funds. The competitive, growing industry of online scholarly publishing reveals the dynamics of knowledge pricing and valuation in contemporary science. Further, given the relatively new institutional challenges of vetting and funding OA research for scholars and scholarly stakeholders, understanding the valuation of OA journals is of particular importance for contemporary scientific policy.

OA scholarly publishing has substantially expanded over the past two decades, occupying complementary and/or competitive niches vis-à-vis established subscription-based journals. Since the early 2000's, there has been a steady increase in OA journals. These increases were driven both via the founding of new journals, as well as the conversion of subscription-based titles to OA. A wide variety of scholars and institutions have founded thousands of OA journals with different philosophies and business models. These heterogeneous niches coupled with the competitive, growing and relatively nascent nature of the OA publishing market underpin substantial variation in journal prices. This wide variation in the OA journal market also enables the analysis of a wide variety of factors that underpin scholarly and economic value in contemporary academic publishing. Social and cultural influences on economic pricing and behaviour are especially germane to the context of science,

where there are social norms discouraging avarice and self-interested behaviour (Merton, 1942).

Hedonic pricing posits that products possess certain attributes or characteristics that are valuable or desirable to consumers (Rosen, 1974). Pricing is influenced by both actual production costs, as well as socio-political forces that influence the valuation of products on both supply and demand sides of the market (Zelizer, 1995; Beckert, 2011). Prices generate needed revenue, but also can function as status signals which influence valuation perceptions of both producers and consumers (Podolny, 2005). Both objective production costs and social sources of value can influence the pricing and valuation of academic outputs, including scholarly journals.

Past research (Björk and Solomon, 2015; Mueller-Langer and Watt, 2018) has shown that journal pricing linked to citation activity. Journals that receive attention, deference and prestige from other publications and scholars are valuable on both supply and demand sides of the publishing market. At the high end of the market, publishers have floated the notion of $25,000 USD APCs for outlets such as *Nature* and *Science* (Pollock, 2018), based on the premise that demand for such prestigious publications is highly inelastic. As suggested by a Springer Nature Publishing executive, "In the end, the price is set by what the market wants to pay for it" (Van Noorden, 2013). Such a philosophy may be at odds with the public good ethos of science. However, scholarly publishing is also a context where science interfaces with business, often creating conflicting institutional logics (Thornton and Ocasio, 1999).

Factors Influencing Journal Valuation

Past research found that journals with higher impact factors and more citations received charge higher APCs (Mueller-Langer and Watt, 2018). In turn, there is a dialectic in the OA journal market, where high-quality journals can charge higher APCs, but the revenue raised from higher prices also underpins increased resources to support legitimate journal quality (Siler et al., 2018). High APCs can fund 'objective' publishing qualities, such as copy editing, professional editors, statistics editors and high-quality typesetting. Subjectively, exclusive journals are selling a prestigious imprimatur – albeit one that publishers may have curated carefully and invested over time – as well as the social signal of affiliation with fellow high-status scientists who publish in such journals. Revenues from high APCs also enable high rejection rates, which can underpin both actual and perceived quality of journals. This raises questions of how much of an APC for a given journal reflects legitimate value, how much is a luxury good (relatively little objective marginal value, augmented by social signalling), and how much is pure profiteering.

The ability and willingness of consumers to pay influences pricing decisions. Accordingly, APCs are often set according to journal or sectoral prestige, as opposed to actual production costs. For example, Elsevier differentially prices journals based on relative funding levels in various academic disciplines (Björk and Solomon, 2015). A 2018 Springer Nature Initial Public Offering on the Frankfurt Stock Exchange candidly promoted the following business strategy for academic journals: *"[W]e intend to employ a price differentiation strategy by tailoring APCs to the discipline and impact factor of the relevant journal[.]…We also aim at increasing APCs by increasing the value we offer to authors through improving the impact factor and reputation of our existing journals"* (p. 99).

As rankings from third parties become increasingly influential in professional fields (Espeland and Stevens, 1997; Espeland and Sauder, 2016), merely being measured is an

important sign of legitimacy for institutions in competitive fields. Even if people view metrics or rankings as unfair or poor measures, they remain important because others take them seriously (Sauder and Espeland, 2009). In science, the Clarivate journal impact factor is the most prominent and influential third-party ranking of journals. Davis (2017) found that when a journal receives its first impact factor, this often leads to increased legitimacy and an influx of new submissions. Receiving and maintaining status endowments like a journal impact factor or DOAJ Seal requires continued legitimacy and conformity to institutionalized criteria. Clarivate annually – sometimes controversially – "de-lists" (removes impact factors) of journals deemed to be engaging in excessive self-citation or exhibiting signs of intellectual balkanization (Davis, 2018). In turn, marshalling the resources – financial and personnel – in order for a journal to attain status endowments, (e.g., impact factor, DOAJ Seal) is an important challenge for publishers and journal stakeholders.

Scholarly publishing involves tensions between economic and scholarly priorities. Publishing is both an economic and scientific activity. Different journals and publishers have different underlying goals and philosophies, which can span the entire continuum between purely academic and purely profit-seeking. This variation in publishing institutions contributes to wide variation in journal pricing. Journals published by commercial publishers tend to be more expensive than those published by not-for-profit organizations. (Bergstrom, 2001). Further, large publishers tend to offer higher-status, more expensive publications than smaller publishers (Björk and Solomon, 2012). Publishers of varying size and status occupy different economic and intellectual niches in the scholarly communication market.

The affluence and level of development of the home countries of academics and their institutions influence scholarly productivity (May, 1997; King, 2004). Inclusion in global scientific networks is conducive to scientific productivity for nations and individual scientists alike (Sugimoto et al., 2017). Both geography and politics influence scholarly collaboration and citation behaviour (Frenken et al., 2009). Scientific journals are institutions via which academic communities can either promote or suppress geographic diversity (Chavarro et al., 2014). Topical priorities in the scholarly corpus are shaped by scientific reward structures, which often devalue or balkanize 'local' concerns in peripheral locations in the global and scientific political economy (Ciarli and Ràfols, 2018). There are also broad concerns that scholarly reward systems tend to overlook and devalue work from peripheral regions and nations (Meneghini et al., 2008). The lowered barriers to entry of online academic publishing creates niches and opportunities for less-wealthy scholars and institutions to contribute to the academic corpus. Some topics and fields of study may have intellectual value to certain communities that are relatively less marketable economically. In turn, the OA publishing market is comprised of many different niches along geographic, linguistic and economic lines.

This article uses a large-scale database of OA journals to examine factors that imbue published science with economic value in the scholarly publishing market.

Methods

Data on current Open Access scholarly journals were acquired from the Directory of Open Access Journals (DOAJ). The DOAJ was founded in 2003 by the non-profit Infrastructure Services for Open Access (IS4OA). The DOAJ is also an index of OA journal legitimacy, as journals must adhere to set criteria to be included. In 2014, stricter quality controls were introduced and 3,776 journals were subsequently culled from the DOAJ (Marchitelli et al., 2017). In 2015, the DOAJ introduced the DOAJ Seal of Approval for Open Access Journals to reward journals on the DOAJ list that adhere to practices deemed particularly meritorious: DOI

usage, submission of metadata, digital archiving, machine-readable licensing, generous Creative Commons licensing, granting authors full copyright. In turn, the DOAJ provides a list of legitimate and distinguished OA journals. The dataset for this study was downloaded from the DOAJ website in December 2018, when the database included 12,100 journals.

Dependent Variables. Total publication costs for authors – the sum of submission and publication fees – is the first dependent variable in the study. Due to the exponential distribution of prices among journals with APCs, the dependent variable is the logarithmic value of the re-centred total cost variable. The second dependent variable is a dummy variable of whether the journal charges APCs and/or submission fees to authors. The majority of journals in the DOAJ dataset are 'free' journals and do not involve direct costs to authors. USD was the most common currency in which publishers levied APCs. For APCs levied in other currencies, world currency exchange rates were used as of December 10, 2018 to convert APCs to USD equivalents.

Independent Variables. The 2017 Clarivate Journal Impact Factor (JIF) values for DOAJ journals were culled from the Web of Science Journal Citation Reports. Out of the 12,100 DOAJ journals, only 1,224 had an official journal impact factor. Due to the exclusivity of the JIF, an additional dummy variable was created denoting whether a journal has a JIF.

Journal language(s) were taken from the DOAJ dataset. Peer review duration, the DOAJ Seal of Excellence award, and first listed academic disciplinary affiliation for journals were also taken from the DOAJ list. World Bank Economic and Geographic regions were coded based on the officially listed location of each journal in the DOAJ dataset.

Publisher type was coded based on the listed affiliation of a journal's main publisher in the DOAJ dataset. Large for-profit publishers were defined as those listed by Larivière et al. (2015) as major oligopolistic publishers – Reed-Elsevier, Wiley-Blackwell, Springer, and Taylor & Francis. Any journal published by those publishers or their subsidiaries was coded as being published by a large for-profit publisher. Small for-profit publishers were operationalized as any for-profit publisher that is not linked to the aforementioned 'oligopolistic' publishers. Any publisher affiliated with a college or university was coded as such. However, if the journal was explicitly published by a university press, this was distinguished separately from those journals published by the university as a whole. Professional associations were coded as publishers with a clear mission to serve members of a certain profession, most commonly academic disciplines.

Results

Table 1 reports tabulations of qualities of journals included in the DOAJ dataset.

Table 1 – Tabulations of DOAJ Journal Characteristics

Journal APC		**DOAJ Seal**	
Free Journal	8,795	DOAJ Seal	1,379
APC Journal	3,305	No DOAJ Seal	10,714
Journal Impact Factor		**World Bank Geographic Region**	
Journal has JIF	1,224	East Asia & Pacific	1,825
Journal w/o JIF	10,876	Europe & Central Asia	6,101
		Latin America & Caribbean	2,310
Journal Language		Middle East & North Africa	594
English Only	5,663	North America	800
Partial English	3,763	South Asia	351
No English	2,674	Sub-Saharan Africa	119
Journal Publisher Organization Type		**World Bank Economic Region**	
Large for-profit publisher	1,517		
Not-for-profit organization	944	High	6,191
University Press	378	Upper-middle	3,876
Professional Association	607	Lower-middle	2,003
Small for-profit	1,546	Low	30
University	5,039		
Uncategorized	1,986	**First Listed Journal Subject**	
		Interdisciplinary	804
		Medicine	2,700
		Social Sciences & Humanities	5,971
		Sciences	2,611

Roughly 73% of DOAJ journal are 'free' and do not charge any submission or publication fees. The remaining journals levy APCs, ranging from $.014 (USD) to $5600 (USD).

Among the 1,224 DOAJ journals with a JIF, JIFs ranged from .0190 to 23.333, with a mean value of 2.283. There was overlap between journals awarded Clarivate journal impact factors and journals awarded the DOAJ Seal of Approval for Open Access Journals. Of the 1,370 journals in the dataset awarded the DOAJ seal, 452 also had a JIF. Accordingly, status endowments in academic publishing are correlated, but do not necessarily completely overlap.

English-only journals comprise roughly 47% of the total DOAJ dataset, while journals with partial English (i.e. include English as one of two or more languages) comprise another 31% (see Appendix for the full tabulated list of publishing languages). Only 22% of journals in the dataset do not publish in English at all.

A variety of different types of institutions publish OA journals. In the DOAJ dataset, universities and colleges are the most common publisher of OA journals, comprising 42% of the total dataset. The 'university' category is distinguished from university presses, which comprised 3% of total journals. Large and small for-profit publishers both comprise 13% of

total journals. Not-for-profit organizations (other than universities and professional associations) comprised 8% of the total, while professional associations accounted for 5%. 16% of total journals were not clearly categorizable based on website analysis.

DOAJ journals are situated in a variety of countries around the world, covering all World Bank geographic and economic regions. Journals published in Europe & Central Asia account for one-half of total journals. Latin America & Caribbean and East Asia & Pacific account for 19% and 15% respectively. Despite the relative prominence of North America in science, the region only accounts for 7% of DOAJ journals. The remaining regions – Middle East & North Africa, South Asia and Sub-Saharan Africa – have relatively smaller presences in the DOAJ dataset. There is also economic stratification in DOAJ representation. High income countries account for 51% of total journals, upper-middle countries account for 32% and lower-middle countries account for 17%. Low income countries only publish 30 total journals listed in the DOAJ dataset, accounting for less than 1% of total journals.

Table 2 reports univariate and multivariate analyses of factors influencing APC prices for DOAJ journals. Due to the exponential distribution of APC price levels, the dependent variable is the logarithm of re-centered APC values. Journals that charge in non-USD currencies were converted based on exchange rates as of December 12, 2018. When considering all 12,100 DOAJ journals, having an official Clarivate journal impact factor is significantly associated with APC levels. As another example of the value of institutional endorsement, journals adorned with the DOAJ Seal also charge significantly higher APCs. Specifically examining the subset of 1,224 articles in the DOAJ dataset with official Clarivate journal impact factors, APC levels were strongly associated with APC levels. In the other models in Table 2, analysis is restricted to the DOAJ journals with a non-zero APC. The English language is significantly associated with APC levels. English-only journals charge significantly more APC than partially-English journals, while journals publishing in one or more languages without any English charge significantly lower APCs.

Table 2 – Regression Analysis of Factors Affecting Prices for APC-based OA Journals

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7	Model 8	Model 9	Model 10	Model 11
Journal Impact Factor											
JIF (yes/no)	3.493*** (.078)										.108*** (.014)
JIF (value)		.437*** (.041)									.139** (.051)
Journal Language											
English only			2.363*** (.060)							1.203*** (.061)	.730*** (.056)
Partially English			[omitted]							[omitted]	[omitted]
No English			-.554*** (.091)							-.203** (.075)	-.315*** (.065)
Publisher Type											
Large for-profit				[omitted]							[omitted]
Not-for-profit organization				-2.049*** (.098)							-1.192*** (.080)
University Press				-1.090*** (.127)							-.496*** (.096)
Professional Association				-1.222*** (.097)							-.578*** (.078)
Small for-profit				-.738*** (.051)							-.548*** (.040)
College/University				-3.197*** (.056)							-1.078*** (.064)
Unclassified				-2.306*** (.068)							-1.018*** (.058)
DOAJ Seal											
DOAJ Seal					1.325*** (.057)						.073 (.038)
World Bank Geographic Region											
East Asia & Pacific						-2.763*** (.096)				-.234*** (.090)	-.052 (.079)
Europe & Central Asia						-.431*** (.082)				-.141* (.062)	-.190*** (.055)
Latin America & Caribbean						-2.165*** (.138)				.140 (.120)	.024 (.108)
Middle East & North Africa						-2.535*** (.140)				-.466*** (.117)	-.301** (.013)
North America						[omitted]				[omitted]	[omitted]
South Asia						-2.360*** (.151)				-.216 (.138)	-420*** (.123)
Sub-Saharan Africa						-2.175*** (.179)				-.235 (.147)	-.022 (.135)
World Bank Economic Region											
High Income							[omitted]			[omitted]	
Low Income							-2.188*** (.606)			-.804 (.567)	-.591 (.486)
Lower-middle Income							-3.015*** (.049)			-2.152*** (.081)	-1.664*** (.076)
Upper-middle Income							-2.153*** (.054)			-1.481*** (.069)	-.959*** (.063)
Peer Review Duration											
Average weeks for peer review								.072*** (.010)			.027*** (.005)
Average weeks for peer review (squared)								-.001*** (.000)			-.001*** (.000)
Journal Subject											
Medicine									[omitted]		[omitted]
Interdisciplinary									-1.747*** (.094)		-.365*** (.059)
Social Sciences & Humanities									-1.917*** (.067)		-.453*** (.044)
Sciences									-.602*** (.060)		-.309*** (.035)
Constant	1.261*** (.025)	3.856*** (.131)	4.180*** (.055)	7.195*** (.035)	5.543*** (.032)	6.933*** (.077)	6.778*** (.022)	5.300*** (.095)	6.693*** (.060)	5.774*** (.083)	6.408*** (.094)
R-squared	.152	.084	.449	.540	.140	.372	.586	.017	.229	.655	.749
N	12,100	1,224	3,305	3,282	3,301	3,305	3,305	3,305	3,300	3,305	3,273

The types of institutions that publish OA journals are also related to APC levels. In our dataset, journals published by large for-profit publishers were the most expensive. Journals published by smaller for-profit publishers were the next most expensive. The least expensive journals were published by colleges, universities and other not-for-profit organizations. Of note

is that journals published by university presses were more expensive than those from universities. University presses and professional associations occupy a 'middle ground' on the pricing continuum between relatively highly priced journals published by for-profit institutions, and less costly journals published by universities and other non-profit organizations.

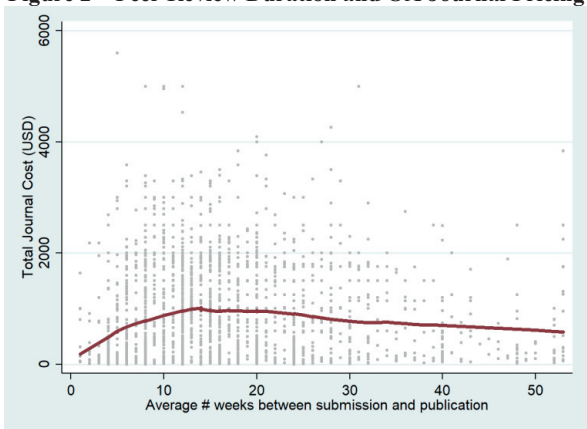
World Bank classifications reveal how geography and political economy are related to OA journal pricing. Journals published in North America were the most expensive, closely followed by the Europe & Central Asia category. The other five geographic regions – Latin America & Caribbean, Sub-Saharan Africa, South Asia, Middle East & North Africa, East Asia & Pacific – all publish journals at significantly lower APC levels than the two wealthiest regions. Analysis of World Bank economic regions reveal that unsurprisingly, journals published in high income countries are significantly more expensive than those published in other economic regions. Moving down the economic hierarchy, journals in upper-middle income countries were more expensive than lower-middle countries. Surprisingly, the few journals published in low income countries were about as expensive as those in middle-income countries, although due to the small number of publishers and journals officially situated in low income countries, this result should be interpreted with caution.

Since geographical, linguistic and economic regions are intertwined, Model 10 combines linguistic and World Bank region data. The effects of English language on journal pricing remain robust, with English-only journals significantly more expensive than partially English journals, and non-English journals significantly less expensive than non-English journals. In contrast to Model 6, when taking linguistic and economic variables into account, journals published in North America were not significantly more expensive than numerous other regions. Pricing in North American journals was not significantly different from journals in Latin America & Caribbean, South Asia or Sub-Saharan Africa. However, even though the relatively small negative coefficient in Model 6 attenuated for Europe & Central Asia, this region remains significantly less expensive than North America in Model 10. This is notable since other regions with much stronger negative coefficients in Model 6 (Latin America & Caribbean, South Asia, Sub-Saharan Africa) attenuated to the point of statistical non-significance in Model 10. Journals in East Asia & Pacific and Middle East & North Africa remain significantly less expensive in Model 10.

Results in the multivariate model in Model 10 for Work Bank Economic Regions reveal similar trends to Model 7. As expected, journals published in high-income countries were most expensive, followed by upper-middle income and lower-middle income countries. However, like with Model 7, there is the somewhat counterintuitive finding that journals published in low income countries are relatively highly priced. In Model 10, pricing for journals in low income countries is not significantly different than those for high income countries. Like with Model 8, this result should be interpreted in light of the relative dearth of articles published in low income countries.

Peer review practices are also related to journal pricing levels. Peer review duration has a curvilinear effect (inverted U-shape) on journal pricing. Model 8 includes both the average weeks for journal peer review, as well as the squared value of that variable. Journals with very fast or very slow peer review processes were relatively less expensive. Figure 2 illustrates the curvilinear effect, suggesting that the 'optimal' peer review duration for OA journal pricing is roughly 12-13 weeks.

Figure 2 – Peer Review Duration and OA Journal Pricing



These results suggest that both very rapid and very slow peer review are conducive to lower journal value, if not also quality.

Model 9 shows how disciplinary orientation influences OA journal pricing levels. Journals in the field of medicine are the most expensive. Journals in the various applied and theoretical sciences were significantly less expensive than the medical journals. Social science and humanities journals were generally even less expensive than journals in the sciences. In short, there is a disciplinary hierarchy in APC levels. Model 11 reports a multivariate model including all major variables in the study. Overall trends regarding OA journal pricing reported in other models in Table 2 remain robust.

Trends for 'free' OA journals were similar to those for APC-based journals. The same variables in Table 2 that were conducive to more conducive to more expensive OA journals (with APCs) were also conducive to a journal being published without APCs. Due to space constraints, this analysis is not included in this draft.

Discussion

The Open Access publishing market is multifaceted with numerous different economic, institutional, social and scientific niches. Even though APC-based and non-APC OA journals occupy different scientific and market niches in contemporary science, similar factors influence both whether a journal charges authors an APC, and price levels for APC-based journals.

The influence of the JIF – both with merely having a JIF, and if so, possessing a higher JIF – underscores the importance and value of citation metrics and third-party evaluation in contemporary science. In turn, publishers and scientists often attempt to bolster or protect the status endowments bestowed by quantitative metrics like the JIF. For many publishers and journals, achieving eminence and status endowments like the journal impact factor require professional and strategic action (Martin, 2016). Some publishers have more resources, knowledge and savvy to achieve prominence and institutionalized esteem for their journals than others. This is one of many mechanisms underpinning cumulative advantage (Merton, 1948) in

science. Publishers and scientists with more resources are both more competitive and able to be reactive to importance metrics in the intellectual and economic markets of science.

The heterogeneity in journal pricing between different types of publishing institutions reveals normative and philosophical conflicts regarding the relationship between economics and science. Scholarly publishing often involves conflicts between scholarly and economic institutional logics. Our results suggest different types of publishing institutions market journals at various points along the continuum between pure economic and pure scientific logics. The scientific institutional logic is underpinned by Mertonian norms, including disinterestedness – the notion that scientists should work solely for the good of science, as opposed to for personal or financial interests (Merton, 1942). Thornton and Ocasio (1999) identified economic and scholarly tensions in scientific publishing, chronicling the historical shift from an editorial logic to a market logic as larger publishing companies consolidated power in the industry starting in the 1970s. Editorial and market logics are not necessarily diametrically opposed; scientific publishing offers a context where the logics can conflict and/or be complementary.

The pricing hierarchy between large for-profit publishers and universities in our results illustrates various institutional logic hybrids involving scientific and economic logics. Journals published by large for-profit publishers were most expensive, followed by smaller for-profit publishers, who perhaps employ similar economic institutional logics to less lucrative economic and institutional niches. Universities and other non-profit organizations published the least expensive journals on the whole, indicative of a strong scientific logic and weaker economic logic. The relatively moderate prices of journals published by university presses and professional associations suggest hybrid economic-scientific institutional logics. Professional associations are often non-profit, but also often rely on journals as a source of institutional revenue. Non-profit organizations are also capable of aggressive rent-seeking, even if such rents are not officially or legally deemed as 'profits.'

Results revealed that journals publishing with the English language occupy relatively more lucrative niches in scholarly publishing market. Over the 19th and 20th centuries, English emerged as the predominant language in science, and now often functions as a *lingua franca* in scholarly communication (Gordin, 2015). In turn, the preeminence of English in academia renders English scholarly journals more economically valuable than journals published in other languages. Evaluative biases in favor of English institutions in science have been identified. For example, Monegon and Paul-Hus (2016) presented evidence suggesting that major scholarly journal databases – such as the Web of Science and Scopus – over-represent English-language journals and tend to exclude non-English journals. Supporting non-English scholarly journals is also often a means of promoting language use and community, particularly for languages vulnerable to being supplanted in professional and social contexts by English. Thus, it makes sense that for many non-English journals, economic logics will be relatively absent. For example, SciElo is a successful database that supports and promotes scientific work in Latin America (Packer, 2009). In the DOAJ dataset, OA journals published via the SciElo database contribute to a disproportionate number of free and low-cost Spanish and Portuguese journals situated in Latin America.

Even after accounting for journal language, the geographic location of publishers was influential on journal pricing. A relative lack of DOAJ journals from less-wealthy countries is notable. Even though the low barriers to entry in OA publishing can enable increased participation in scholarly publishing from traditionally excluded groups and regions (Suber, 2012), economic and geographic stratification remain in contemporary OA publishing. Lower income countries are less likely to publish DOAJ journals, as well as the World Bank

geographic regions of Middle East & North Africa, South Asia and Sub-Saharan Africa. In the case of pay journals, the author-pays model of APC-based OA publishing appears to currently be more accessible to scholars situated in wealthier countries and institutions (Siler et al., 2018). If publishers set prices based on the willingness or ability of scholars and institutions to pay APCs, that is another mechanism supporting higher APCs in wealthier contexts. The overrepresentation of Latin America in the DOAJ database suggests the importance of strong publishing institutions, especially in less-lucrative niches that may not attract large for-profit publishers. However, supporting strong publishing institutions also requires economic and scholarly resources that not all regions or countries may possess.

The curvilinear results for peer review duration and journal pricing also reveal the intellectual and economic niches of downmarket – if not 'predatory' – scholarly journals. Journals with very rapid peer review are relatively less expensive. Likewise, OA journals with relatively long peer review duration are also less expensive. In the case of very rapid peer review, this could often be reflective of haphazard – or non-existent – quality control. Journals that cannot compete on quality may instead attempt to compete on speed, which may be particularly attractive to naïve or unscrupulous scholars. Relatively lengthy peer review processes can also be evidence of journal unprofessionalism and/or insufficient resources to process and peer review manuscripts expediently. Given that the culture of OA publishing is based in part on leveraging the efficiency and immediacy of online dissemination, longer peer review durations may be a much less legitimate or tolerated practice vis-à-vis established print-based journals. More broadly, there are concerns about the financial incentive structure of APC-based OA, where publishers are remunerated for published articles, and article rejections yield costs but no revenues. In turn, APC-based OA publishing can incentivize less-rigorous peer review and lower quality (Jeon and Rochet, 2010; Gans, 2017: 56). At worst, journals that offer very fast peer review as part are 'predatory' journals, which offer haphazard peer review, facilitating the exchange of money for an 'easy' publication.

Examining the relationship between journal pricing and academic disciplines reveals an 'economic' hierarchy of the sciences. Journals in medicine and the hard sciences are most expensive, while social science and humanities journals are least expensive. This suggests that journals are often priced in part based on the ability of potential consumers to pay, as more pecunious disciplines are expected to tolerate higher APCs. Exploitation of willingness-to-pay in academic publishing may be seen as morally questionable given norms against unbridled profit-seeking in science (Merton, 1948). More generally, profit-maximizing behaviour – or strict adherence to economic logics above all other considerations – is often seen as culturally inappropriate or undesirable in many contexts deemed of societal importance (Zelizer, 1995). The variety of economic niches in scholarly publishing, as well as the various factors that imbue academic journals with economic value reveal the complex interplay between economic and scientific logics in contemporary science.

References

Beckert, Jens. 2011. "Where do prices come from? Sociological Approaches to Price Formation." *Socio-Economic Review*, 9: 757-786.

Bergstrom, Theodore C. 2001. "Free Labor for Costly Journals?" *Journal of Economic Perspectives*, 15(4): 183-198.

Björk, Bo-Christer and David Solomon. 2012. "Pricing principles used by scholarly open access publishers." *Learned Publishing*, 25(2): 132-137.

Björk, Bo-Christer and David Solomon. 2015. "Article processing charges in OA journals: relationship between price and quality." *Scientometrics*, 103(2): 373-385.

Chavarro, Diego, Puay Tang, and Ismael Ràfols. 2014. "Interdisciplinarity and research on local issues: evidence from a developing country." *Research Evaluation*, 23(3): 195-209.
Davis, Phil. 2017. "PeerJ Membership Model and The Paradox Of The Loyal Customer." https://scholarlykitchen.sspnet.org/2017/05/08/peerj-membership-model-paradox-loyal-customer/
Davis, Phil. 2018. "Tipping the Scales: Is Impact Factor Suppression Biased Against Small Fields?" https://scholarlykitchen.sspnet.org/2018/10/08/tipping-the-scales-is-impact-factor-suppression-biased-against-small-fields/
Espeland, Wendy N., and Michael Sauder. 2007. "Rankings and Reactivity: How Public Measures Recreate Social Worlds." *American Journal of Sociology*, 113(1): 1-40.
Espeland, Wendy N., and Mitchell L. Stevens. 1998. "Commensuration as a social process." *Annual Review of Sociology*, 24: 313-343.
Frenken, Koen, Shoerd Hardeman, Jarno Hoekman. 2009. "Spatial scientometrics: Towards a cumulative research program." *Journal of Informetrics*, 3(3): 222-232.
Gans, Joshua S. 2017. *Scholarly Publishing and its Discontents*. Toronto: Core Economic Research.
Gordin, Michael D. 2015. *Scientific Babel: How Science Was Done Before and After Global English*. Chicago: University of Chicago Press.
Jeon, Doh-Shin, and Jean-Charles Rochet. 2010. "The Pricing of Academic Journals: A Two-Sided Market Perspective." *American Economic Journal: Microeconomics*, 2(2): 222-255.
Larivière, Vincent, Stefanie Haustein, and Philippe Mongeon. 2015. "The Oligopoly of Academic Publishers in the Digital Era." *PLOS ONE*, 10(6): 1-15.
Marchitelli, Andrea, Paola Galimberti, Andrea Bollini and Dominic Mitchell. 2017. "Helping journals to improve their publishing standards: a data analysis of DOAJ new criteria effects." *JLIS.it*, 8(1): 1-21.
Martin, Ben R. 2016. "Editors' JIF-boosting stratagems – Which are appropriate and which not?" *Research Policy*, 45(1): 1-7.
Meneghini, Rogerio, Abel L. Packer and Lilian Nassi-Calò. 2008. "Articles by Latin American Authors in Prestigious Journals Have Fewer Citations." *PLOS ONE*, 3(11):e3804.
Merton, Robert K. 1942. *The Sociology of Science*. Chicago: University of Chicago Press.
Merton, Robert K. 1968. The Matthew Effect in Science. *Science*, 159: 56-63.
Monegon, Philippe, and Adèle Paul-Hus. 2016. "The journal coverage of Web of Science and Scopus: a comparative analysis." *Scientometrics*, 106(1): 213-228.
Mueller-Langer, Frank, and Richard Watt. 2018. "How many more cites is a $3,000 Open Access fee buying you? Empirical evidence from a natural experiment." *Economic Inquir*y, 56(2): 931-954.
Packer, Abel L. 2009. "The SciELO Open Access: A Gold Way from the South." *Canadian Journal of Higher Education*, 39(3): 111-126.
Pollock, Dan. 2018. "News and Views: Open Access Charges." https://deltathink.com/news-views-open-access-article-processing-charges/
Rosen, Sherwin. 1974. "Hedonic Prices and Implicit Markets: Product Differentiation in Pure Competition." *Journal of Political Economy*, 82(1): 34-55.
Sauder, Michael, and Wendy N. Espeland. 2009. "The Discipline of Rankings: Tight Coupling and Organizational Change." *American Sociological Review*, 74(1): 63-82.
Siler, Kyle, Stefanie Haustein, Elise Smith, Vincent Larivière, and Juan Pablo Alperin. 2018. "Authorial and institutional stratification in open access publishing: The case of global health research." *PeerJ*, 6: e4269.
Suber, Peter. 2012. *Open Access*. Cambridge: MIT Press.
Sugimoto, Cassidy R., Nicolas Robinson-Garcia, Dakota S. Murray, Alfredo Yegros-Yegros, Rodrigo Costas and Vincent Larivière. 2017. "Scientists have the most impact when they're free to move." *Nature*, 550: Oct 5., 29-31.
Thornton, Patricia H., and William Ocasio. 1999. "Institutional Logics and the Historical Contingency of Power in Organizations: Executive Succession in the Higher Education Publishing Industry, 1958-1990." *American Journal of Sociology*, 105(3): 801-843.
Van Noorden, Richard. 2013. "Open access: The true cost of science publishing." *Nature*, 495: 426-429.
Zelizer, Viviana A. 1995. *The Social Meaning of Money*. Princeton: Princeton University Press.

Making sense of global collaboration dynamics: Developing a methodological framework to study (dis)similarities between country disciplinary profiles and choice of collaboration partners

Nicolas Robinson-Garcia[1], Richard Woolley[2] and Rodrigo Costas[3]

[1] *elrobinster@gmail.com*
Delft Institute of Applied Mathematics, TU Delft, Netherlands
INGENIO (CSIC-UPV), Universitat Politècnica de València, Spain

[2] *ricwoo@ingenio.upv.es*
INGENIO (CSIC-UPV), Universitat Politècnica de València, Spain

[3] *rcostas@cwts.leidenuniv.nl*
CWTS, Leiden University, Netherlands
DST-NRF Centre of Excellence in Scientometrics and Science, Technology and Innovation Policy, Stellenbosch University, South Africa

Abstract

This paper presents a novel methodological framework by which the effects of globalization on international collaboration can be studied and understood. Using the cosine similarity of the disciplinary and partner profiles of countries by collaboration types it is possible to analyse the effects of globalization and the costs and benefits of an increasing global networked research system.

Introduction

The growth of scientific collaboration and co-authorship in the last century has drawn the attention of bibliometricians and sociologists of science for quite some time now. In the words of Cole (1973), "it is generally recognized today that science in fact develops within a community of interacting scientists" (p. 1). The seminal works by Beaver & Rosen (1978) and Crane (1972) have led to a profusion of studies analysing collaboration in many ways, including motivations for collaboration (Katz & Martin, 1997), collaboration strategies (Bozeman & Corley, 2004), collaboration structure (Newman, 2001) or benefits derived from international collaboration in terms of scientific impact (Bote, Olmeda-Gómez, & Moya-Anegón, 2013), among others.

The increase of international collaboration experienced in the last decades has been described as an effect of the globalization of the research system, leading to an increase of international exchange and flows of mobile scholars (Czaika & Orazbayev, 2018), and facilitating collaboration between distant countries (Waltman, Tijssen, & Eck, 2011). It has been argued that such changes could decrease national disparities by 'flattening' the world and reducing zones of inclusion and exclusion in a global scientific collaboration network (Saxenian, 2005; Woolley, Robinson-Garcia, & Costas, 2017). But empirical studies suggest that international collaboration networks are organized hierarchically (Wagner, Whetsell, & Leydesdorff, 2017), showing an inverse relation between the growth of the share of collaborative papers and the geographical spread of these networks (Wagner, 2005). These findings along with the fact that networked or multilateral collaboration is increasing overtime (see table 1) raise many questions with regard to how countries are adapting to this new landscape, their integration into these global networks and the potential 'costs' or 'benefits' this integration may bring together.

A fundamental element of understanding of the dynamics of integration in global scientific networks is based on the work of the German sociologist Georg Simmel (1950: 135), who famously recognised the effect of introducing a third actor into a social context: "there is, in addition to the direct relation between A and B, for instance, their indirect one, which is derived from their common relation to C". Simmel took this expansion from the dyad to the triad as the basis for a crucial piece of understanding, that indirect relationships are essential to the formation and cohesion of the groups and sub-groups that characterize the social world. Social network analysis has since built on this foundational distinction to investigate social structures (Burt 1992) and study how the roles and positioning of actors in social networks effects the distribution of power (Brass and Burkhardt 1992) and the diffusion of information (Granovetter 1973).

We take up this distinction between dyads and triads as the basis for our understanding of two modes of international collaboration in science: partnerships and networks. International research partnerships and networks are both constructed from the same fundamental element, what network theorists call "mutual dyads" in which the relationship between two actors is based on mutual recognition and reciprocity (Wasserman and Faust 1994: 511). We only focus on mutual dyads (and not on 'directional' dyads) because we tend to assume that collaboration relationships in science are based on direct interpersonal relationships characterized by varying degrees of both trust and conflict (Shrum et al. 2001). However, whereas we understand international research partnerships to be based on a single mutual dyad, we understand international research networks as being based on a combination of two or more mutual dyads. This network mode can be illustrated most simply in a triadic form (Figure 0).

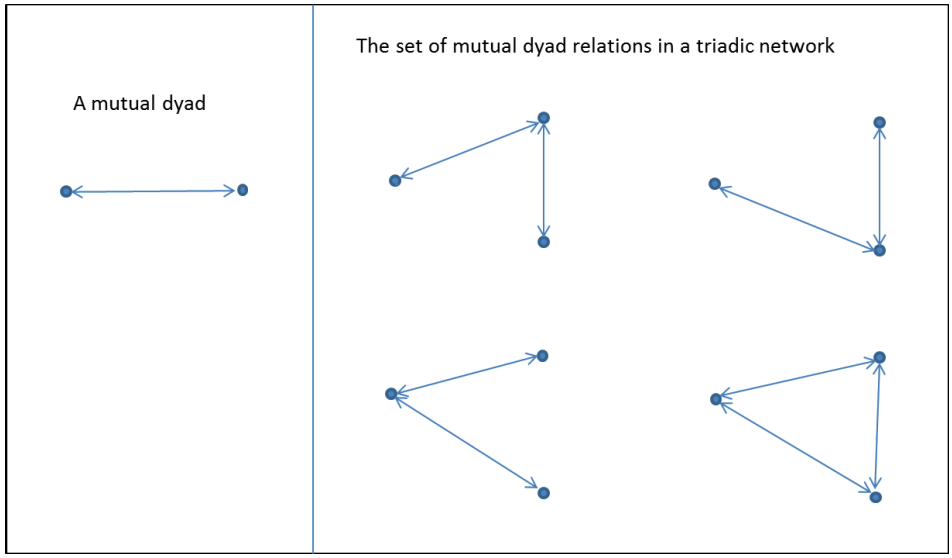


Figure 0. Mutual dyads in triadic relationships

A triadic collaboration network can be made up of four potential sets of mutual dyads (Fig 1). In three of the four possible triadic networks the relationship between two of the nodes only *only* indirect, that is, mediated via the third node. International research collaborations can thus be characterised theoretically as containing "structural holes" (Burt 1992), where the

triangular relationship is not closed on all sides by a mutual dyad. We apply this logic to international scientific collaboration to assume that 1) whilst all collaboration relationships are based on mutual reciprocity and exchange based on trust relations; 2) it is not necessarily the case that all collaborators must have a direct relationship with all other collaborators. While network theorists argue that "transitive" networks, in which the triad is 'closed' and all elements are connected are more stable and likely to be more durable (Burt 1992; Wasserman and Faust 1994), we simply do not know whether this is the case in relation to scientific collaborations or not. Rather, a very significant limitation of network analyses based on bibliometric data is that the qualities of networks remain obscured.

The distinction between bilateral international research collaboration (BIRC) and multilateral international research collaboration (MIRC) is thus important because discussions of global co-authorship networks implicitly assume equality of connectedness within the network, assuming that a homogeneous set of mutual dyadic relations exist between each co-author of a scientific paper. In other words, a default assumption of homogeneity in sets of bibliometric relationships slides into an assumption of equality in sets of sociological relationships. While this is not intended as a criticism of what bibliometric-based network analyses can do, it is intended as a reminder of the limitations of bibliometric network analyses in relation to assuming 'connectedness' as a homogeneous form of social organisation of knowledge production - and then characterising so-called 'global networks' on this basis. Although we are not able to understand whether the multilateral networks in our data are transitive (closed) or not, this theoretical approach will nevertheless have significant implications for how we interpret our results.

Table 1. Growth rate by WorldBank regions for the 1980-2017 period for Bilateral International Research Collaboration (BIRC) and Multilateral International Research Collaboration (MIRC) collaboration. Data source: Web of Science

	BIRC	MIRC
East Asia & Pacific	11.4%	15.8%
Europe & Central Asia	7.4%	12.6%
Latin America & Caribbean	9.5%	15.2%
Middle East & North Africa	8.8%	14.3%
North America	7.4%	12.6%
South Asia	9.8%	16.1%
Sub-Saharan Africa	8.3%	13.3%

In this study, we propose a new methodological framework by which the dynamics of globalization through international scientific collaboration can be studied and start to be better understood. For this, we suggest comparing countries' disciplinary profiles and choice of collaboration partner (i.e. the collaboration partner profile) by publication type. We distinguish the following publication types: 1) domestic publication (publications authored by one or more scholars affiliated to a single country), 2) bilateral international collaboration (publications co-authored by scholars affiliated to two different countries) and 3) participating in a collaboration network (publications co-authored by scholars affiliated to three or more countries). We base our empirical work on the model proposed by (De Lange & Glänzel, 1997; Glänzel & De Lange, 1997) who previously distinguished between no collaboration, BIRC and MIRC. We hypothesize that countries' international profile will differ from their domestic profile, with this difference being greater for their MIRC profile. In other words, as countries become drawn into multilateral international networks they move away from the

focus (topics) that characterise domestic knowledge production. We suggest that this methodology has the potential to better inform studies focused on specific countries or regions to better understand not only how integrated they are in global networks, but also the potential effects or factors which can explain their specific situations.

Data and methods

Data collection

We collected all publications for all countries for the 2008-2017 period. This data was gathered from the CWTS-enhanced version of the Web of Science. Country information was extracted from each publication, normalized and linked to the World Bank regions classification. For each publication we counted the number of different countries to which authors were affiliated and created a new collaboration type field in which we distinguished between BIRC, MIRC and domestic publications. Disciplinary profiles of countries are created based on the distribution of their publications among the Web of Science subject categories classification. We produced four disciplinary profiles for each country: 1) based on their domestic output, 2) based on their international output (co-authored with at least another country), 3) based on their BIRC output (co-authored just with one other country) and 4) based on their MIRC output (co-authored with at least two other countries). This distinction is based on the model used by (De Lange & Glänzel, 1997; Glänzel & De Lange, 1997), although we acknowledge that this approach could be extended by including further divisions. For instance, Adams & Gurney (2018) suggest that publications authored by 20 or more countries should be treated differently due to their special nature.

Methodological approach

Here we propose measuring similarity of a countries' domestic disciplinary profile with their BIRC and MIRC disciplinary profiles by using the cosine similarity (Salton & McGill, 1986). Cosine similarity is usually employed in bibliometric studies when analyzing co-occurrence data such as co-citation networks or co-citation networks (e.g., Aman, 2018; Wagner, 2005; Yan & Ding, 2012).

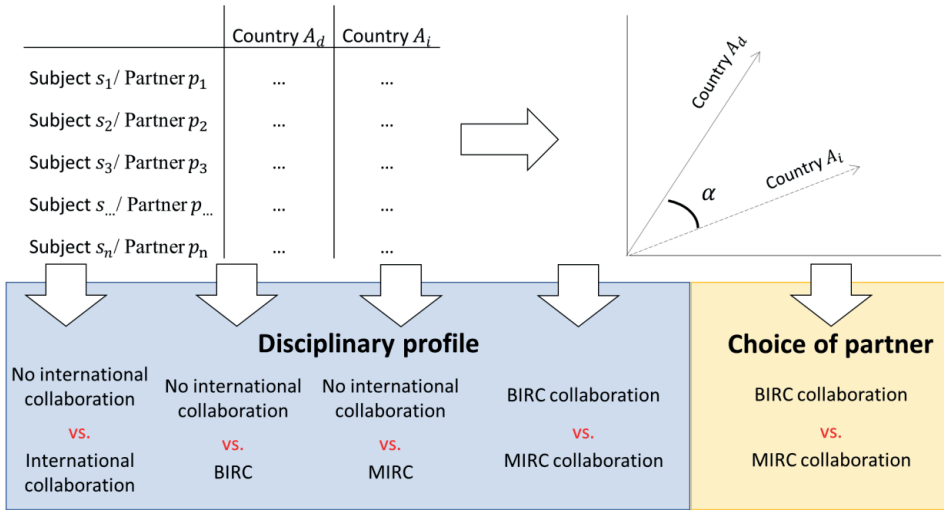


Figure 1. Calculation of five similarity variables per country for their disciplinary and collaboration partners profiles.

Figure 1 presents a schematic representation of our approach, depicting the profile distribution of subjects (s_n) or partners (p_n), for domestic publications (A_{dom}) and international publications (A_{int}), and how the cosine analysis is applied for the disciplinary profile on the one hand, and the choice of partner on the other hand. Let A_d be the disciplinary profile of country A when publishing domestically where $A_{dom} = \{s_1, s_2, \dots, s_n\}$ being s_n the number of publications n in subject s. Let A_{int} be the disciplinary profile of country A when publishing with international collaboration. The similarity between two disciplinary profiles is defined as:

$$SIM(A_{dom}, A_{int}) = \cos(\alpha) = \frac{A_{dom}\, A_{int}}{|A_{dom}||A_{int}|}$$

Where a value of 0 indicates no similarity between profiles and 1 indicates that both profiles are identical. This same procedure can then be applied to all combinations of publication types to find (dis)similarities between disciplinary profiles. Furthermore, it can also be calculated to identify (dis)similarities on countries' partner of choice distribution. To this end, we compute five similarity indicators for each country as shown in Figure 1. While four of them relate to disciplinary similarity, one relates to similarity on collaboration partner. The rationale of this is that one would expect that a high disciplinary (dis)similarity between BIRC and MIRC profiles, would lead to a high (dis)similarity in the distribution of publications by collaboration partners between BIRC and MIRC profiles.

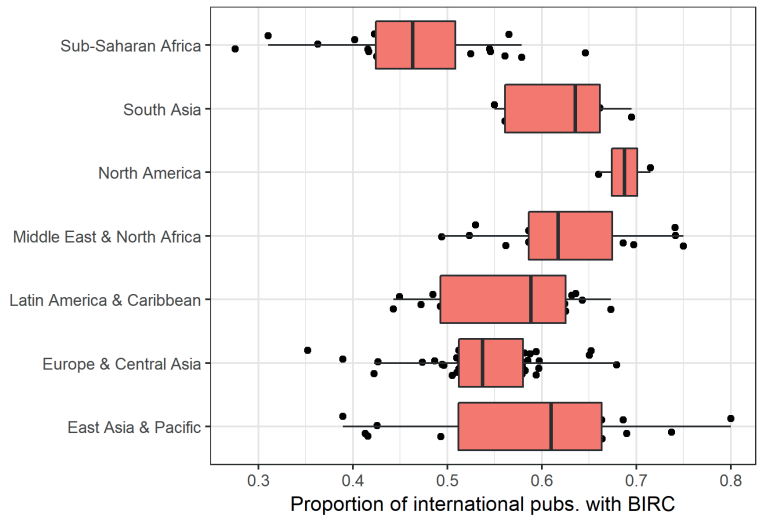


Figure 2. Proportion of publications with BIRC by WorldBank regions for the 2008-2017 period

Results

In this section we show some results of a global analysis of international collaboration. Figure 2 shows the proportion of BIRC for each region in the world. In general, the majority of regions' output in international collaboration is bilateral with the exception of Sub-Saharan Africa (on median, 46% of their output is bilateral). On the other extreme we find North America (69%) or South Asia (64%). Overall we find an important dispersion among

countries within all regions, with no significant differences between regions (with the exception of Sub-Saharan Africa and North America.

Figure 3 illustrates how cosine similarities between the different profiles of a country disaggregated by publication type, can be used to better understand disciplinary differences between countries. The figure highlights similarities and differences between international collaboration and domestic knowledge production (Figure 3A), between collaborating bilaterally and collaborating multilaterally (Figure 3B), and between choice of partner countries when collaborating bilaterally and collaborating multilaterally (Figure 3C). In these figures, countries in blue are those which exhibit a similarity of profiles above the world average, while those in orange exhibit a similarity below world average. Overall, we observe that differences are more acute between the disciplinary profiles of countries when collaborating internationally versus not collaborating than they are between BIRC and MIRC. Furthermore, we observe that, with the exception of the United States on choice of partner (Figure 3C), countries with a similarity below average tend to belong to Eastern Europe, Africa and, to some extent, Asia.

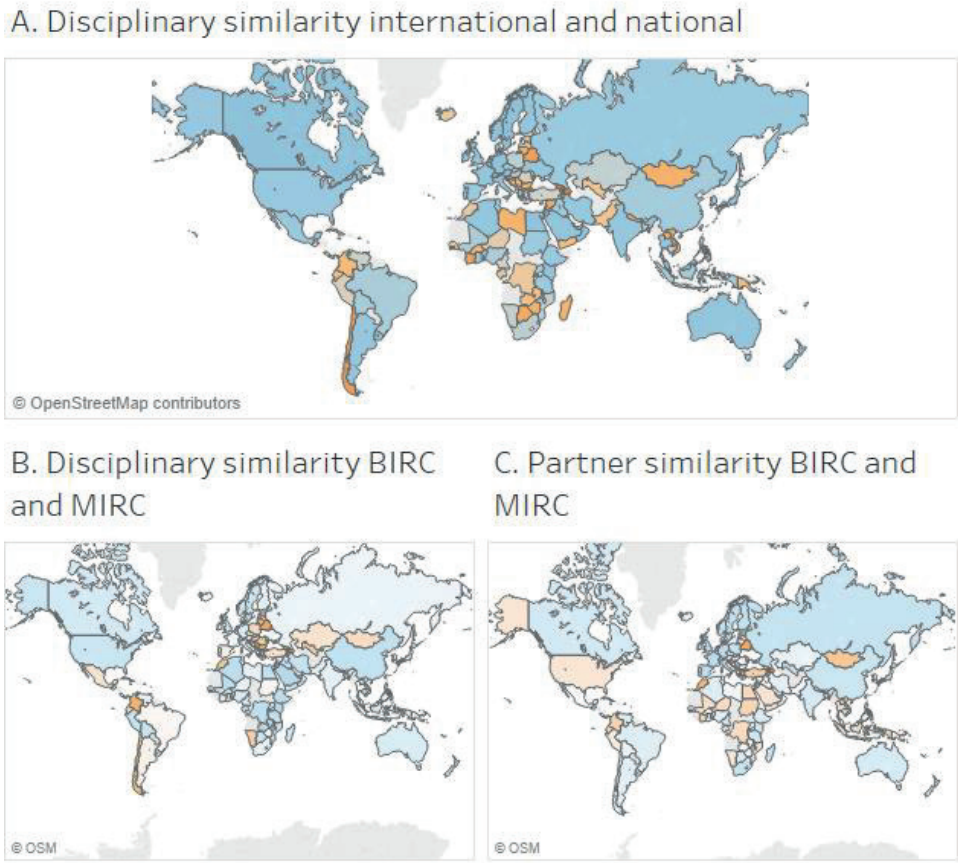


Figure 3. Cosine similarities for each country according to their A) disciplinary domestic and international profiles, B) disciplinary BIRC and MIRC profiles, and C) collaboration partners' BIRC and MIRC profiles. Blue intensity indicates similarity above world average; Orange intensity indicates similarity below world average. (Colors only in online version)

When looking into which type of international collaboration deviates more from the disciplinary profile of countries' domestic publications (Figure 4), we observe that MIRC does appear to be more dissimilar than BIRC profiles. But we can also see that these differences vary greatly between countries and regions. For instance, we observe large differences for Latin America & Caribbean, and specifically for Chile and Colombia. Also Eastern European and Central Asian countries seem to show a larger difference than Western and Central European countries. Indeed, we observe that while the similarity of European and Central Asian countries between their BIRC and national disciplinary profiles is high, it is also relatively homogeneous between countries, while there are larger disparities when comparing the similarity of their MIRC and national disciplinary profiles.

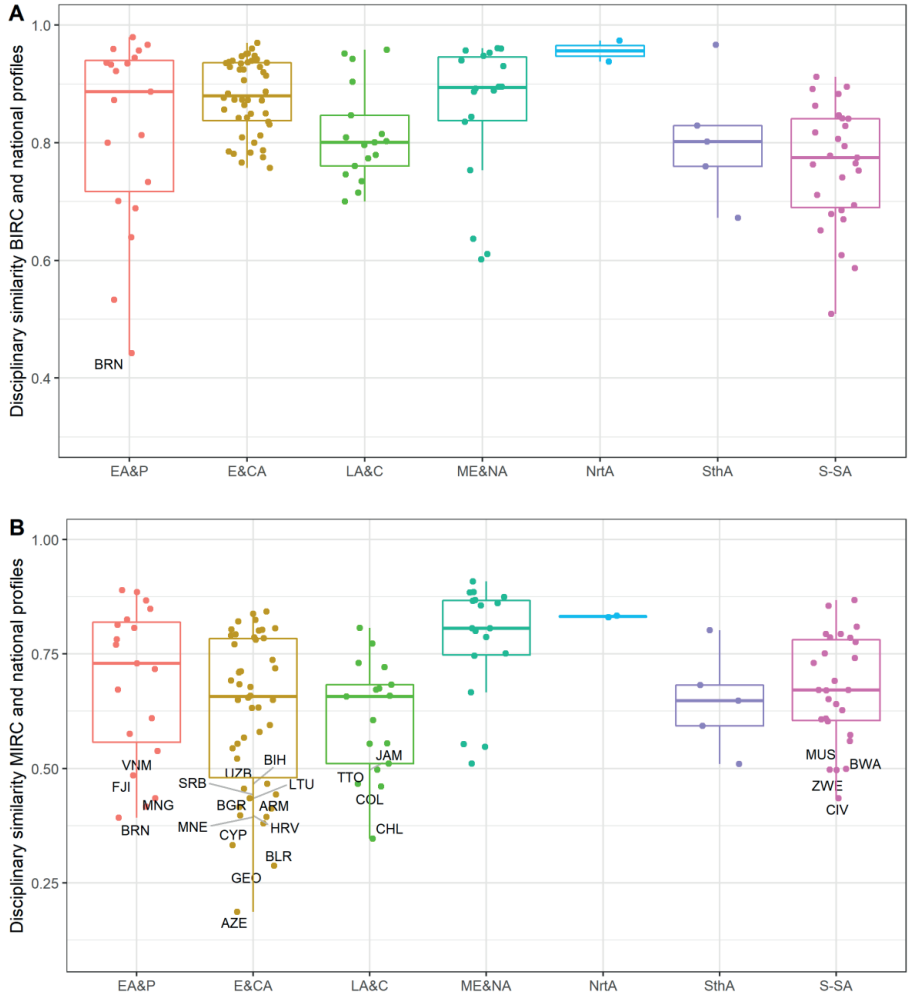


Figure 4. Boxplots showing similarity of countries by region for A) their BIRC versus national and B) their MIRC versus national disciplinary profile. Countries with value under 0.5 are highlighted

Figure 5 and 6 illustrate further ways by which these indicators could be employed to better interpret specific situations. Figure 5 shows in the y-axis the similarity of the disciplinary profile of countries from East Asia & Pacific when collaborating internationally versus when not doing so (domestic profile). The x-axis shows the proportion of publications internationally co-authored they produce. We observe that the least productive countries are not only the ones which show a higher dependency on international collaboration, but many of these smaller countries exhibit greater disciplinary differences when collaborating internationally from the domestic profile.

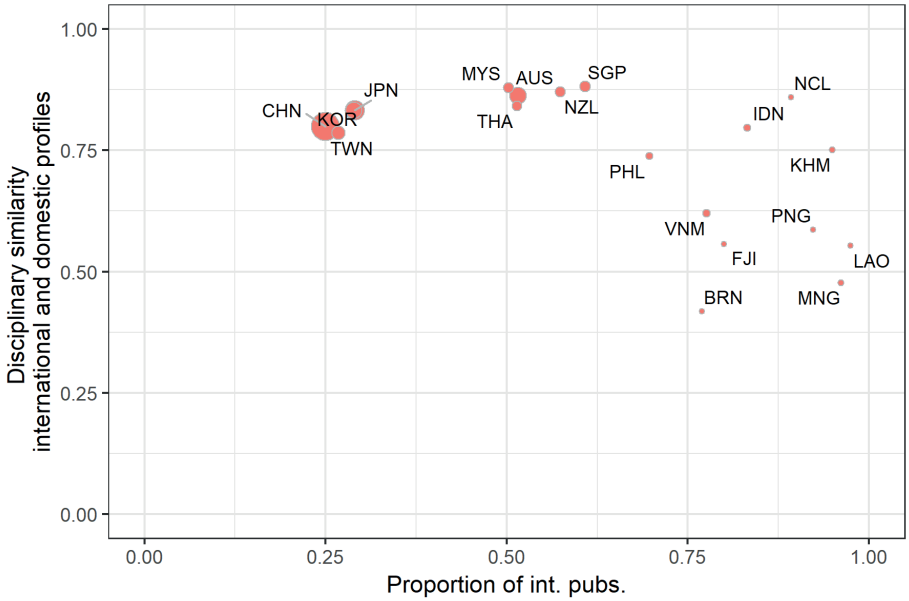


Figure 5. Scatterplot comparing the disciplinary similarity of the international versus domestic profile of each country and the proportion of publications internationally co-authored for countries in East Asia & Pacific. Size of dots indicates total number of publications

In Figure 6 we look into both Europe & Central Asia and Sub-Saharan Africa and compare similarities between countries' BIRC and MIRC disciplinary profiles with their BIRC and MIRC choice of partner. Countries that exhibit lower disciplinary similarities also exhibit lower similarities on choice of partners in Europe and Central, however this is not always the case for Sub-Saharan African countries. Western European countries show a higher similarity for both indicators, while interestingly South Africa shows a high partner similarity but a lower level of disciplinary similarity when comparing BIRC and MIRC

Concluding remarks

In this paper we propose deconstructing countries' publication profile based on the collaboration type of their output. We suggest that by using similarity measures and comparing countries' collaboration profiles both in terms of their distribution of disciplines and choice of partner, we take an important step toward the development of a better understanding of the international partnerships and networks that shape the dynamics of globalising science.

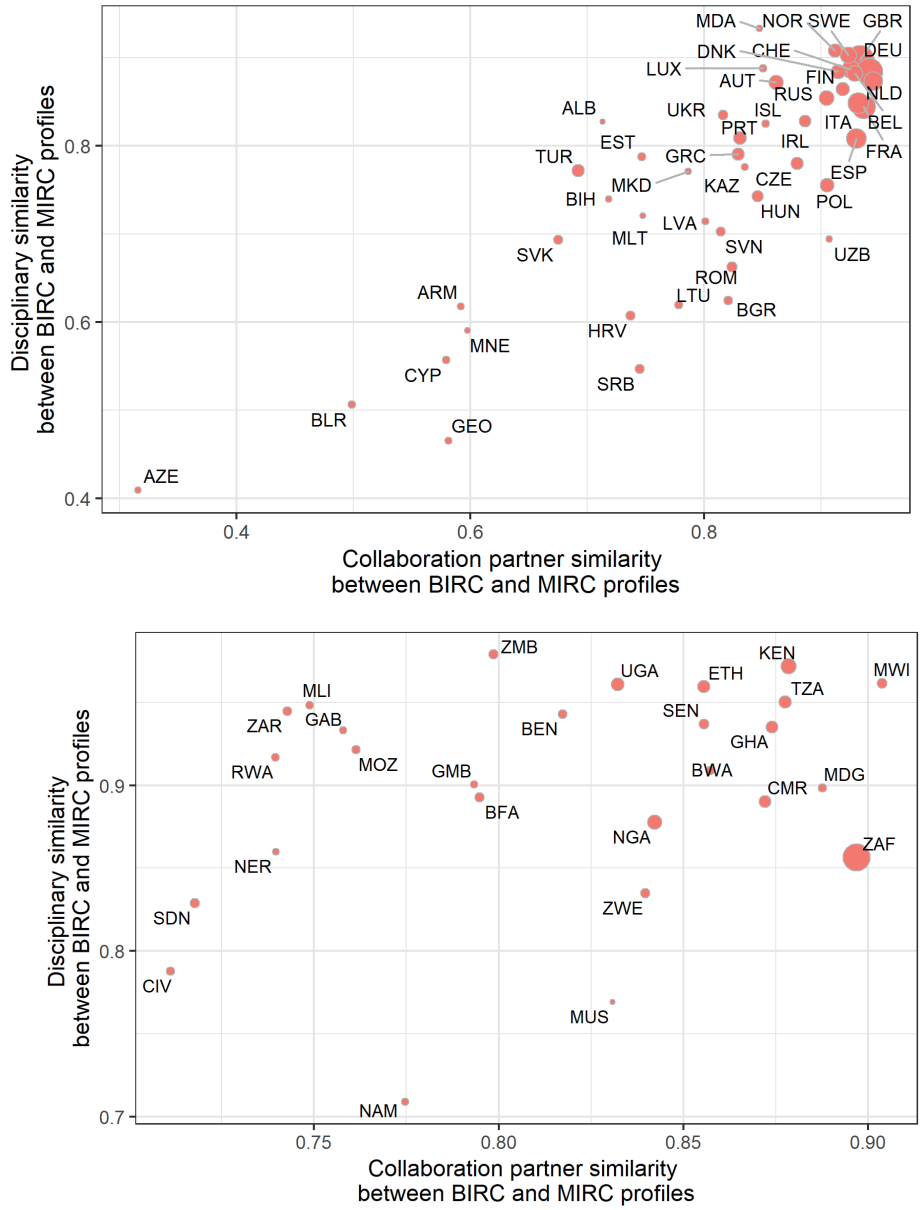


Figure 6. Scatterplot comparing the disciplinary similarity of the BIRC versus MIRC and the collaboration partner's similarity of the BIRC versus MIRC profile for each country in Europe & Central Asia (above) and Sub-Saharan Africa (below). Size of dots indicates number of internationally co-authored publications

We have suggested the use of the cosine as a measure of similarity to establish a set of 'internal' comparisons at the national level, in the sense that we are always comparing different portions of a country's output. We have shown how this method can help us to

understand the connections between what a country's scientists are working on (topic, discipline) and who they are working on it with (international partners). We have also shown the efficacy of the method for comparing between countries and regions on these dimensions. Further methodological development will follow to identify other types of benchmarking of international collaboration that can be designed from this foundation.

Our theoretical distinction between international research partnerships (BIRC) and international networks (MIRC) can also form the basis for future efforts to better understand relationships between formalised patterns of scientific co-authorship and the social integration underpinning scientific collaborations. Further technical advance and the use of complementary methodologies will be required in this respect. But we consider the present methodological innovation to be a constructive first step in this direction.

Acknowledgments

Nicolas Robinson-Garcia is a Marie Sklodowska Curie experienced researcher in the LEaDing Fellows COFUND programme sponsored by the European Commission. This research is partially funded by the South African DST-NRF Centre of Excellence in Scientometrics and Science, Technology and Innovation Policy (SciSTIP).

References

Adams, J., & Gurney, K. A. (2018). Bilateral and Multilateral Coauthorship and Citation Impact: Patterns in UK and US International Collaboration. *Frontiers in Research Metrics and Analytics*, *3*. Recuperado mayo 4, 2018, a partir de https://www.frontiersin.org/articles/10.3389/frma.2018.00012/full

Aman, V. (2018). A new bibliometric approach to measure knowledge transfer of internationally mobile scientists. *Scientometrics*, *117*(1), 227-247.

Beaver, D. deB, & Rosen, R. (1978). Studies in scientific collaboration. *Scientometrics*, *1*(1), 65-84.

Bote, V. P. G., Olmeda-Gómez, C., & Moya-Anegón, F. de. (2013). Quantifying the benefits of international scientific collaboration. *Journal of the American Society for Information Science and Technology*, *64*(2), 392-404.

Bozeman, B., & Corley, E. (2004). Scientists' collaboration strategies: implications for scientific and technical human capital. *Research Policy*, Scientific and Technical Human Capital: Science Careers and Networks as Knowledge Assets, *33*(4), 599-616.

Brass D.J. & Burkhardt, M.E. (1992). Centrality and power in organizations. In: Nohria N and Eccles RG (eds) *Networks and Organizations: Structure, Form, and Action*. Cambridge, MA: Harvard Business School Press, pp. 191–215.

Burt, R. (1992). *Structural Holes: The Social Structure of Competition*. Cambridge, MA: Harvard University Press.

Chinchilla-Rodríguez, Z., Bu, Y., Robinson-García, N., Costas, R., & Sugimoto, C. R. (2018). Travel bans and scientific mobility: utility of asymmetry and affinity indexes to inform science policy. *Scientometrics*, 1-22.

Cole, J.R. (1973). *Social stratification in science.* Chicago ; London: University of Chicago Press.

Crane, D. (1972). *Invisible colleges: Diffusion of knowledge in scientific communities*. Chicago: University of Chicago Press.

Czaika, M., & Orazbayev, S. (2018). The globalisation of scientific mobility, 1970–2014. *Applied Geography*, *96*, 1-10.

De Lange, C., & Glänzel, W. (1997). Modelling and measuring multilateral co-authorship in international scientific collaboration. Part I. Development of a new model using a series expansion approach. *Scientometrics*, *40*(3), 593–604.

Glänzel, W., & De Lange, C. (1997). Modelling and measuring multilateral co-authorship in international scientific collaboration. Part II. A comparative study on the extent and change of international scientific collaboration links. *Scientometrics*, *40*(3), 605–626.

Granovetter MS (1973) The strength of weak ties. *American Journal of Sociology* 78(6): 1360–1380.

Katz, J. S., & Martin, B. R. (1997). What is research collaboration? *Research Policy*, *26*(1), 1-18.

Newman, M. E. J. (2001). The structure of scientific collaboration networks. *Proceedings of the National Academy of Sciences*, *98*(2), 404-409.

Salton, G., & McGill, M. J. (1986). Introduction to modern information retrieval.

Saxenian, A. (2005). From Brain Drain to Brain Circulation: Transnational Communities and Regional Upgrading in India and China. *Studies in Comparative International Development*, *40*(2), 35-61.

Shrum, W., Compalov, I:, & Genuth, J. (2001) Trust, Conflict and Performance in Scientific Collaborations. *Social Studies of Science* 31(5): 681-730. doi: 10.1177/030631201031005002

Simmel, G. (1950). *The Sociology of George Simmel*. K Wolf (Ed.), Glencoe, Illinois: Free Press.

Wagner, C. S. (2005). Six case studies of international collaboration in science. *Scientometrics*, *62*(1), 3-26.

Wagner, C. S., Whetsell, T. A., & Leydesdorff, L. (2017). Growth of international collaboration in science: revisiting six specialties. *Scientometrics*, *110*(3), 1633-1652.

Waltman, L., Tijssen, R. J. W., & Eck, N. J. van. (2011). Globalisation of science in kilometres. *Journal of Informetrics*, *5*(4), 574-582.

Wasserman, S. & Faust, K. (1994) *Social Network Analysis: Methods and Applications*. Cambridge: CUP:

Woolley, R., Robinson-Garcia, N., & Costas, R. (2017). Global research collaboration: Networks and partners in South East Asia. *arXiv:1712.06513 [cs]*. Recuperado marzo 28, 2018, a partir de http://arxiv.org/abs/1712.06513

Yan, E., & Ding, Y. (2012). Scholarly Network Similarities: How Bibliographic Coupling Networks, Citation Networks, Cocitation Networks, Topical Networks, Coauthorship Networks, and Coword Networks Relate to Each Other. *Journal of the American Society for Information Science and Technology*, *63*(7), 1313-1326.

Technological specialization of cities: a new patent-based approach and evidence from Russia

Ekaterina Streltsova[1] and Gleb Kuzmin[2]

[1] *kstreltsova@hse.ru*
National Research University Higher School of Economics, Institute for Statistical Studies and Economics of Knowledge, Myasnitskaya st., 11-443, Moscow (Russia)

[2] *gkuzmin@hse.ru*
National Research University Higher School of Economics, Institute for Statistical Studies and Economics of Knowledge, Myasnitskaya st., 11-445, Moscow (Russia)

Abstract

In the pursuit of smart specialization and evidence-based policy, nowadays researchers and analysts need to depart from country-level evaluations and concentrate on a more targeted analysis of regions, agglomerations and smaller geographical units. This might be a challenge as some types of data necessary to fulfil this task are aggregated on the national or regional levels only. The paper looks at patent statistics which is one of the kind. It offers a new combinatory approach to collect and aggregate data for cities and discusses the results of this methodology testing. The study identified the cities which are the most important contributors to the technological development of Russia. Their technological specialization and patent portfolios were assessed, which helped better understanding of the competitive advantages and potential of these agglomerations for further technological growth.

The paper will be of interest not only for practitioners who make decisions at the regional or municipal levels, but also for researchers working in the field of regional economics, economic geography, economics of science, technology and innovation. A particular value the research might bring to the analysts from the countries where city-level patent statistics is unavailable or challenging to obtain.

Introduction

Being centers of science, technology, and production, large cities are traditionally considered as the main "growth points" of national economies (Boschma, Balland & Kogler, 2014; Jacobs, 1969; Glaeser et al., 1992; O'hUallachain, 1999). Such agglomerations tend to have higher economic productivity and innovation activity, since they concentrate resources and create conditions for the free exchange of ideas between individuals, organizations, industries (Andersson, Quigley & Wilhelmsson, 2005; Balland, 2015a; Carlino & Kerr, 2014; Jacobs, 1969; 1984; Jaffe, Trajtenberg & Henderson, 1993). As a result, large cities are in the subject area of various disciplines, including spatial and regional economics, economic geography, economics of science, technology and innovation. In addition to theory development, they all solve practical problems, studying among other things the technological potential and technological specialization of agglomerations.

Russia is rich of large cities: 15 of them have a population of more than one million inhabitants. Similar to many other countries, Russian cities are the national key drivers of the technological development. Most technologies are created here, which is demonstrated by patent activity indicators. In 2017, about 30% of all domestic patent applications in Russia were filed by residents of the 'two capitals' – Moscow and St. Petersburg, which are the only cities in the country with available patent statistics (Rospatent, 2018). The technological specialization of these and other cities, their technological priorities, and orientation towards the domestic or global market are issues that have not been given sufficient attention in the academic literature. The reason is the lack of necessary data or the technical difficulty of obtaining it: the available statistics are aggregated to the level of the regions of the Russian Federation, while other sources (for example, patent databases) allow information search only at the national level (by

country of inventor or applicant). The situation is typical for many countries: the analysis of city specialization and technological portfolio is as demanded as it is difficult to carry it out.
This fact was an initial motivation for the study which elaborates and tests a new methodology to assess the technological capabilities of the cities that are the Russian leaders in the development of new technologies. The research task has important practical implications, since the method allows one to determine current thematic priorities and potential of cities, and to predict their future technological growth and vulnerability in the case of technological crises. The relevance of such work is confirmed by the number of studies devoted to assessing the impact of specialization on the technological diversification of cities, and the dynamics of their inventive activity.

Technological specialization and sustainability of cities

Technological specialization is analyzed by economists at various levels – for individual types of organizations (Dachs, Mahlich & Zahradnik, 2007; Pattel & Pavitt, 1991), industries (Ha et al., 2015), regions and countries (Archibugi & Pianta, 1992; Ejermo, 2005; Pianta & Meliciani, 1996). Less often, cities are also the object of such research. The study by Cortright & Mayer (2001), which is one of the pioneer papers in this field, discusses the specialization of 14 US cities which are the centers for high technology. With an analysis of the data on employment, patent activity, and venture capital, the authors found that despite the general orientation of these cities to the development of high-tech industries, each of them has its own and narrow technological specialization. For example, Atlanta was focused on creating databases, Boston – on computer technology, medical equipment, software development, Denver – on data storage technologies, telecommunications, etc. The dynamics of the specialization of US cities and the variability of their trajectories are also discussed in the studies by Rigby (2015) and Kogler, Heimeriks & Leydesdorf (2013). Some research in this area was conducted in other countries, for example, Germany (Vlckova, Kasprikova & Vlckova, 2018), China (Xia & Hu, 2014).
A cross-country comparison was made in a collective study (Kogler, Heimeriks & Leydesdorf, 2018), which compared the patent portfolios of 20 large cities in five countries (China, France, Israel, the Netherlands, and the US). It revealed significant differences in the technological specialization of agglomerations, including those from one country. The results are of great practical importance, as they clearly demonstrate the impossibility of applying a one-fits-all approach to managing technological and innovative development of cities.
The close attention of researchers to technological specialization at all levels is explained by its potentially high economic impact: understanding the current situation allows one to identify the competitive advantages of an organization, region or country, and to determine their position in the structure of regional, national or global technology markets (Giannitsis & Kager, 2009). If it is followed by sound and effective management decisions, this knowledge might help to transform technological specialization into technological leadership. Such an analysis is particularly relevant in economic crises or out-of-crisis funding cuts, when the need to determine investment priorities is particularly acute.
The assessment of specialization is useful not only for understanding the achieved capabilities, but also for predicting the future technological progress. Researchers emphasize that this process is path dependent and its trajectories are limited to a set of technologies that are already being developed in the country, region, city (Cantwell & Vertova, 2004; Strumsky, Lobo & van der Leeuw, 2012). This is evidenced by the study of Boshma and his colleagues (Boschma, Balland & Kogler, 2014), based on an analysis of the patent activity of 366 US cities in 1981–2010. It revealed that a new technology is more likely to emerge and develop successfully in a city if it is related to the existing technological portfolio. In contrast, technologies that are radical in relation to existing areas, more at risk of gradual regression and extinction.

These results are confirmed by Rigby (2015), who also studied the dynamics of various technologies in major US cities and its dependence on their current technological specialization. He found that in most cities, competencies are concentrated around a limited set of related technologies which determine the technological development in the future. "The core" is characterized by a high degree of inertia – radical changes in the technological specialization of cities are rare and slow.

Understanding the specialization of regions and cities can also be used to predict their susceptibility to technological crises – long periods of decline in inventive and, as a result, patent activity due to various external or internal factors. A number of studies on regional / urban technological resilience show that cities which are technologically diversified (i.e. developing a whole range of unrelated technologies) are less likely to experience technological crises, recover faster after them, and tend to more effective 'technological renewal'. Highly specialized cities, on the contrary, tend to experience technological declines more intensively as measured by patent activity, more often and longer (Balland, Rigby & Boschma, 2015b; Boschma, 2015).

The results of these studies are in favor of assessing the technological portfolio of cities, and demonstrate the relevance of developing new approaches and metrics to perform this task. First, it allows defining the "core" of competences of agglomerations which are often the drivers of national technological development. Secondly, it helps to predict the trajectories of further development of technologies in cities, to understand their potential to diversify. Finally, an analysis of the scope of specialization makes it possible – albeit in the form of hypotheses – to assess the probability of the onset of technological crises in cities and their potential to recover. Russia is a suitable case to test a new methodology designed to perform this task as its technological progress is dependent on the activity of large cities. A lack of such studies in the country also confirms the relevance of the task.

Data and method

The assessment of technological specialization is traditionally based on the analysis of patent activity, its thematic structure and dynamics (Grilliches, 1990; Gokhberg, 2003). In most technological areas, obtaining a patent is the most popular method of protecting R&D results, therefore patent documents are an important source of information on new technologies (Boschma, Balland, & Kogler, 2014). They contain detailed information about the inventor and assignee, the country of origin and patent office, the date of filing or granting, etc., which allows solving numerous research problems. Moreover, each patent document indicates the technological domain to which the patented subject matter relates (Fleming & Sorenson, 2001). In most countries, including Russia, the International Patent Classification (IPC) is used for these purposes (https://www.wipo.int/classifications/ipc/en/). Using the codes of this or alternative classifications, researchers evaluate the thematic structure of patents and the dynamics of technology development. For this study, the distribution of patent documents by technology areas is made in accordance with the Technology Concordance Table (Schmoch, 2008), which is an adapter between the IPC and 35 technological domains, including computer technology and digital communication, pharmaceuticals and biotechnology, microstructural and nanotechnology.

Patent analysis procedures are generally standardized and widely known, but despite this, an assessment of the patent activity of cities and their technological specialization is a non-trivial task. Available open access and commercial patent databases are not designed for an objective and detailed analysis of these territorial units. First, they do not allow automated searching for patent documents by region or city of origin (to our knowledge, the only exceptions are US and – to some extent – European Patent Office which provides statistics on the level of NUTS3

regions). Secondly, the assignee address is often missing, especially in patent documents which are originally non-English. As a part of this study, a method of working with patents was elaborated, which allowed us to overcome these and other technical limitations. We believe it might help researchers from the countries with no city-level patent statistics available to better understand the technological capabilities of agglomerations which are often the key drivers of national economy. The algorithm to develop an empirical base for the research included several steps.

First, *a registry of resident patent applications filed in Russia* was developed. These documents and metadata could not be imported from the Russian patent office directly as the system does not filter the applications or patent grants by the status of the applicant (resident / non-resident), nor includes the indication of the technological domains each invention refers to, nor allows unlimited downloading for further processing offline. As a result, this task was performed using the PatStat Global database, which aggregates information from most patent offices around the world. The period from 2008 to 2016 was chosen for analysis, which allowed us to evaluate the emerging trends and to avoid random, short-term fluctuations in patent activity. Initially, a 10-year period was considered (from 2008 to 2017), but the study found that information for 2017 was only partially published in the database used (a significant time lag in updating data is a common feature and limitation of all databases containing primary patent information). As a result, it was decided to shorten the period. It also allowed for the calculation of indicators for equal, three-year periods.

Due to technical difficulties in obtaining such data, the foreign patent applications of Russian applicants were not considered. However, in the past 5 years, the share of such applications has averaged 14%, so their exclusion from the sample does not significantly affect the results of the analysis.

At this initial phase, more than 180,000 patent applications were uploaded to our own PostgreSQL database.

As a second step, *a registry of applicants* was formed, i.e. a list of unique names of organizations (for legal entities) and personal names (for individuals). This was done by exporting information from the corresponding field of previously downloaded patent applications. After the deletion of erroneous lines, the registry volume was 55,000 units. In order to obtain objective results, two applicants (individuals) with extremely high patent activity rates were excluded from the register, otherwise they would distort the results not only in a particular city, but in the whole country. (One of them tends to apply for hundreds of patents in food chemistry annually. Due to his activity, Russia takes the second position globally on the number of patent applications in this field, while its place in the consolidated rating is 11th only).

Thirdly, the registry of applicants was supplemented with *home / companies' addresses*. For this purpose, an automated search for this information was undertaken in the Russian patent database.

For each unique applicant, a search was made for one (in the case of several – random) application by document number, then the address specified in the document was added to our own database and distributed to all applications of this applicant. This approach potentially has some limitations, for example:

- all applications of one applicant are allocated to one place of residence, although it may change;
- the postal address specified in the document was automatically considered as the address of residency of the applicant, while it may be an address of a patent attorney or organization providing such services;

- it ignores the possibility of complete coincidence of the names of applicants living in different regions. However, the likelihood of such risk is generally low: random check of the data did not reveal a single such case.

Finally, *the distribution of patent applications among cities* were made. This task was performed automatically by processing the postal codes contained in the address field of the document (with a use of the registry of the Russian Post). Each patent application was assigned to a specific city.

After the empirical basis was developed, the cities were ranked by the number of domestic applications in 2008-2016. For the cities that occupied high positions, key indicators were calculated:

- the total number of patent applications filed in Russia (by year);
- average annual growth of patent activity;
- the share of 35 technological domains in the patent applications filed by the residents of a certain city (the proportion of technology). PatStat Global uses fractional count to attribute inventions and related patent documents to specific technological domains. As this database was the primary source of information, the study also follows this approach, which avoids multiple counts of inventions which refer simultaneously to several technological domains. The is used to calculate this indicator and those listed further.
- the city's share in the total number of patent applications filed in Russia and related to each of the 35 technological domains (the proportion of the city);
- concentration indices – C_5 and C_10, which are calculated as the sum of the weights of the 5 and 10 largest technological domains in the city's portfolio, and characterize its level of specialization or diversification;
- revealed technological advantage index (RTA), which is one of the most traditional metrics to analyze technological specialization (Gokhberg, 2003; Khramova et al., 2013). It is calculated by comparing the structure of patent applications filed by residents of a particular city with the general structure of resident applications in Russia. In this study, we considered as technological specialization only domains with RTA exceeding 1.1, i.e. presented in the city's patent portfolio much stronger than the national average.

The results of the testing of the offered methodology along with the key evidence of the analysis are presented below. Primary data and calculations are open for further research at the link: http://bit.ly/2HfVPzd.

Research results

The leading positions in the ranking of Russian cities by the number of patent applications are expectedly occupied by Moscow and, with a large lag, St. Petersburg. Nine other cities – Voronezh, Ufa, Kazan, Novosibirsk, Yekaterinburg, Krasnodar, Perm, Samara and Tomsk – constitute a homogeneous group: they are significantly inferior to the leaders, but ahead of the cities included in the second ten of this rating. These cities (Figure 1) strongly contribute to the technological development of the country: they account for more than half of all patent applications filed by residents in Russia. Nevertheless, they are not the largest agglomerations in the country: Voronezh, Ufa, Krasnodar and Perm are outside the top ten in terms of population, and Tomsk occupies only 28th position in this rating (Rosstat, http://www.gks.ru). Unfortunately, at the moment, due to limitations of statistical data available in Russia for city level, it is not possible to analyze the correlation of patent activity of cities and indicators characterizing the resources for R&D (R&D funding or personnel, etc.). Nevertheless, such a

Figure 1. The leading cities in patent activity in Russia: 2014-2016

simple comparison of the number of patented technologies and the population size demonstrates that cities use their resources in different ways. In general, these results are consistent with the findings of previous studies, which showed that the greatest number of inventions are created in large agglomerations, although individual small cities can also succeed in this process (O'hUallachain, 1999). As a rule, an important factor in this case is the presence of a successful university or federal research center.

Over the past decade, the leaders have not changed significantly, which indicates the absence of sharp spikes in patent activity in other cities and the stability of the previously established system of distribution of forces in the Russian intellectual property market.

The technological "core" of the cities: specialization versus diversification

The theoretical part of the article stated that the more extensive the technological specialization of the city, the higher its ability to develop new technologies and recover from technological crises. Our analysis revealed that from this point of view, two Russian cities have the greatest potential – Moscow and Novosibirsk, whose technological portfolios are distinguished by a high degree of diversification (Figure 2). In the capital, only about one third of all patented

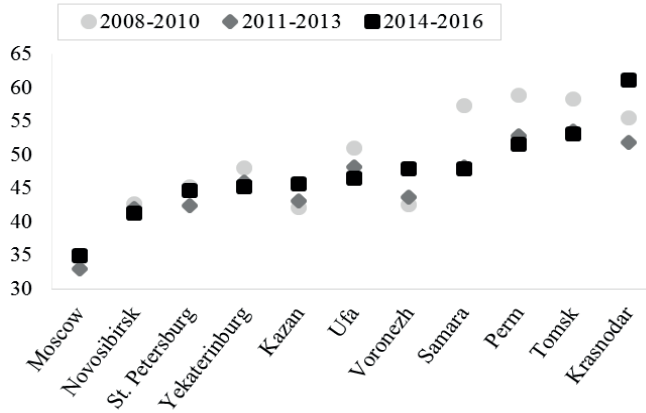


Figure 2. Dynamics of the concentration index of the cities (C_5)

inventions fall on the five largest technological domains.
The highest concentration indices are in Perm (51.6%), Tomsk (53.1%) and Krasnodar (61.1%), which indicates a clear dominance in their portfolio of a limited number of technologies. Moreover, in Krasnodar the concentration index (C_5) has increased significantly in 2008-2016, which goes against the general tendency for the Russian technological cities to gradually diversify.

Technological specialization
The key results of the research are presented in Table 1: the color highlights the areas that are within the technological specialization of the cities. By reading the table horizontally, one can identify cities with a potential for the development of certain technologies, and having sufficient basic conditions for this. Reading the table in columns, one can get an idea of the technological portfolios of the cities and the areas that each of them specializes in. Providing the reader with the opportunity to independently decide which task is his/her priority, we note here only the most important results which demonstrate the relevance of the study and its approach to patent analysis.
The study revealed that three categories of technological cities can be distinguished in accordance with the areas of their specialization. To identify them, we considered the adapter classification of IPC and NACE (Van Looy et al., 2014) and the OECD Taxonomy of Economic Activities Based on R&D Intensity (Galindo-Rueda, Verger, 2016).
The first category is the cities with the dominance of high tech, and focused on new, innovative areas. In Russia, this category includes Moscow, St. Petersburg and Tomsk.
Moscow is specialized in a number of domains related to ICT, and the city generates most new information technologies in Russia. It is the place of residency of large IT companies with high patent activity: Kaspersky Lab, Yandex, ABBYY. In computer technology, Moscow accounts for about 50% of all patent applications filed by residents in Russia in 2014–2016.

Table 1. Technological specialization of the cities: 2014 – 2016

		Technological domains	*Moscow*	*St. Petersburg*	*Voronezh*	*Ufa*	*Kazan*	*Novosibirsk*	*Yekaterinburg*	*Krasnodar*	*Perm*	*Samara*	*Tomsk*
Electrical engineering	1	Electrical machinery, apparatus, energy	*	*		*		1.50	1.63	*		1.14	2.27
	2	Audio-visual technology	1.15	1.45					1.30				
	3	Telecommunications		1.67	2.37								1.98
	4	Digital communication	1.16	3.42								1.21	
	5	Basic communication processes		1.30	3.86								
	6	Computer technology	1.85	1.52									
	7	IT methods for management	1.73	1.16			2.71		1.85			1.15	
	8	Semiconductors	1.44	1.22	1.23			1.67					1.55

		Technological domains	*Moscow*	*St. Petersburg*	*Voronezh*	*Ufa*	*Kazan*	*Novosibirsk*	*Yekaterinburg*	*Krasnodar*	*Perm*	*Samara*	*Tomsk*
Instruments	9	Optics		1.63			2.59	4.09					
	10	Measurement	*	1.77	*	*	*	1.43	*	*	*	*	1.73
	11	Analysis of bio materials							1.19		2.66	1.27	2.33
	12	Control		1.33			1.27					1.12	
	13	Medical technology	1.12	1.31	*	1.10	1.11	*	1.26	*	*	1.78	1.60
Chemistry	14	Organic fine chemistry				6.44	1.32	1.96	1.14	1.35	2.35		1.24
	15	Biotechnology	1.87					2.18					
	16	Pharmaceuticals	1.10	*		1.94	1.11	1.25	*	*	*	*	1.71
	17	Macromolecular chemistry, polymers	1.11		1.18	1.32	2.02						1.43
	18	Food chemistry	*		4.01				*	6.21			
	19	Basic materials chemistry	*			3.04	2.92			1.72	1.41		1.13
	20	Materials, metallurgy	*	*			*	*	2.66		1.58	1.43	1.67
	21	Surface technology, coating			1.14	1.53	1.28	1.14	1.39		1.22		
	22	Microstructural and nano-tech	1.16					2.01	1.69		1.32		1.57
	23	Chemical engineering	*	*	*	1.87	*	1.96	*	1.95			1.15
	24	Environmental technology				1.96							
Mechanical engineering	25	Handling		1.52	1.46				1.65		1.26		
	26	Machine tools							1.13			1.34	
	27	Engines, pumps, turbines	*	*	1.33	*	1.80		*		2.27	2.07	
	28	Textile and paper machines		1.37			1.26				1.47		
	29	Other special machines	*	*	1.47		*	*		2.07	*	*	
	30	Thermal processes and apparatus			2.02	1.29		1.97	1.16				
	31	Mechanical elements	*			*	*					1.26	
	32	Transport	*	1.45			*		*			1.14	
Other fields	33	Furniture, games			2.30				3.33		1.78	1.13	
	34	Other consumer goods	1.65							1.42			
	35	Civil engineering	*	*	*	1.13	1.42	*	*		2.67	1.16	*

Note: Colored cells – areas of technological specialization of the city (RTA is indicated in the cells)
* – "large" areas that are not included in the technological specialization of the city (the proportion of the domain in the total number of patent applications of the city exceeds the average value)

Moscow specializes in several other high-tech domains – for example, biotechnology, microstructural and nanotechnology. In the first case, the city accounts for almost 50% of all patent applications, in the second – about 30%. Semiconductors, basic materials chemistry, surface technology and coating are areas that have also characterized its specialization over the past decade. In general, the technological portfolio of this leading city is highly diversified, it seems to have strong capabilities in many domains, which increases the chances of the emergence of new technologies here.

St. Petersburg, the second largest city of Russia, specializes in the technologies in almost the entire spectrum of electrical engineering, including audio-visual technology, telecommunications, computer technology. Digital communication is the domain where St. Petersburg has achieved the most prominent success (it accounts for 26% of all patent applications).

The technological specialization of *Tomsk* includes measurement (this is the largest domains in its patent portfolio), organic fine chemistry, pharmaceuticals, microstructural and nanotechnology. None of the two last sections of the classifier ("Mechanical Engineering" and "Others") is in the specialization area of the city.

The second category of the technological cities are those mostly focused on the development of less knowledge-intensive, low technologies. For the Russian case, it includes Krasnodar and Perm.

Krasnodar specializes in a whole block of chemical areas – food chemistry, basic materials chemistry, chemical engineering, etc. In some of them, it demonstrates high patent activity: Krasnodar accounts for more than 13% of domestic patent applications in food chemistry (2014-2016). More than 25% of the inventions patented by the residents of this city belong to this domain. The field of technological specialization of *Perm* includes a number of areas related to chemical and mechanical engineering: engines, pumps, turbines; materials, metallurgy; paper and textile machinery; basic materials chemistry, etc. It is worth mentioning that some high technologies are being successfully developed in Perm: organic fine chemistry, and microstructural and nanotechnology. This city can be considered a border case: at the moment, more traditional areas dominate in its portfolio, but the dynamics of several technological domains suggests that the current model of specialization may change over time.

And finally, the third category is comprised of cities specializing in a whole range of technologies, regardless of their technical level or knowledge intensity. Among them, are Voronezh, Ufa, Kazan, Novosibirsk, Yekaterinburg and Samara.

Voronezh specializes in diverse domains: from digital communication to other special machines. The diversification of this city is also demonstrated by the areas in which it makes the most significant contribution. Food chemistry is leading here (Voronezh accounts for 11.5% resident patent applications, while on average it is just 3.1%) and basic communication processes (11.0%).

Ufa is clearly focused on the chemical technologies: this block accounts for half of all the patent applications filed by the residents of this city. Ufa makes a special contribution to the development of organic fine chemistry in the country (14% of the patent applications), which is primarily ensured by the activity of the key developer in this area – Institute of Petrochemistry and Catalysis, Russian Academy of Sciences (RAS).

The technological specialization of *Kazan* includes semiconductors, macromolecular chemistry and polymers, organic fine chemistry, basic materials chemistry. Optics occupies an outstanding position in the technological portfolio of *Novosibirsk* – about 9% of Russian patented inventions in this area in 2014-2016. The city has considerable potential for the development of biotechnology: it traditionally takes the third position in the rating of the Russian cities by a number of patent applications in this domain. The Institute of Chemical

Biology and Fundamental Medicine, RAS, is the most active applicant of the city in this area and secures its high position.
The analysis of concentration indices, mentioned above, demonstrates that the transformations of the technological portfolios of *Ekaterinburg* and *Samara* in the last decade have been characterized by gradual diversification. As a result, today both cities specialize in new technologies within a whole range of domains, both high and low tech.

Conclusion

In Russia, as in most countries, the drivers of technological development are large cities. Due to the availability and geographical proximity of resources – financial, human, technical – this process proceeds here most intensively, which is reflected in the indicators of patent activity. Over the past decade, the rating of the leading cities has remained virtually unchanged. On the one hand, such stability might indicate the absence of significant breakthroughs from other players in the Russian technology market, on the other hand, might be regarded as an indicator of the high potential and sustainability of cities previously included in the rating. The strategies and capabilities of these cities are extremely diversified. This might be an obvious and rather predictable outcome, if consider differences in their location and access to natural resources, well-being and the presence of large scientific and educational centers, along with other factors. Nevertheless, the practical relevance of the study lies not only in the empirical confirmation of this intuitively formulated hypothesis, but also in the comprehensive analysis of technological specialization and patent portfolios of the Russian leading cities. Its results can be used in decision-making to introduce new measures to support technological development at regional and even municipal levels.
The study allowed us to identify technologies being equally developed by all the Russian leading cities. They might be considered as a reliable basis for further technological growth of the country. Such domains include measurement. Three cities out of our 11-item list specialize in these technologies (St. Petersburg, Novosibirsk and Tomsk), but the others are also actively involved: for almost each of them, it is among the largest domains in terms of the number of patent applications. Due to the involvement and performance of developers from various regions, in measurement patents, Russia is the 8th largest globally, although in all areas in total – takes only the 11th position. A similar situation is in medical technology, civil engineering and – in recent years – pharmaceuticals.
Despite some limitations of this study, discussed in the methodological section, it suggests a convenient and already tested approach to patent-related research and sets the basis for further analysis of technological development in cities. The study sets the floor for the investigation of the factors influencing patent activity of agglomerations and specificity of their technological profile. To do so, a city-level statistical data on R&D funding, R&D personnel, equipment, etc. is necessary, which is a challenge still to be solved.

Acknowledgments

The article was prepared within the framework of the Basic Research Program at the National Research University Higher School of Economics (HSE) and supported within the framework of a subsidy by the Russian Academic Excellence Project “5-100”.

References

Andersson, R., Quigley, J.M. & Wilhelmsson, M. (2005). Agglomeration and the special distribution of creativity. *Regional Science*, 84 (3), 445-464.

Archibugi, D & Pianta, M. (1992). Specialization and size of technological activities in industrial countries: The analysis of patent data. *Research Policy*, 21 (1), 79-93.

Balland, P.-A., Boschma, R. & Frenken, K. (2015a). Proximity and Innovation: From Statics to Dynamics. *Regional Studies*, 49 (6), 907-920.
Balland, P.-A., Rigby, D. & Boschma, R. (2015b). The technological resilience of US cities. *Cambridge Journal of Regions, Economy and Society*, 9, 167-184.
Boschma, R. (2015). Towards an Evolutionary Perspective on Regional Resilience. *Regional Studies*, 49 (5), 733-751.
Boschma, R., Balland, P.-A. & Kogler, D.F. (2014). Relatedness and technological change in cities: the rise and fall of technological knowledge in US metropolitan areas from 1981 to 2010. *Industrial and Corporate Change*, 24 (1), 223-250.
Cantwell, J. & Vertova, G. (2004). Historical evolution of technological diversification. *Research Policy*, 33, 511-529.
Carlino, G. & Kerr, W. (2014). Agglomeration and Innovation. *Working Paper 15-007. Harvard Business School*. Retrieved November 15, 2018 from: http://real.wharton.upenn.edu/~duranton/Duranton_Papers/Handbook/Agglomeration_and_innovation.pdf.
Cortright, J. & Mayer, H. (2001). High Tech Specialization: A Comparison of High Technology Centers. *Center on Urban and Metropolitan Policy*. Retrieved November 15, 2018 from: https://www.brookings.edu/wp-content/uploads/2016/06/specialization.pdf.
Dachs, B., Mahlich, J. & Zahradnik, G. (2007). The Technological Competencies of Korea's Firms: A Patent Analysis. In Mahlich, J. & Pascha, W. (Eds.), *Innovation and Technology in Korea: Challenges of a Newly Advanced Economy* (pp. 127-146). NY: Physica-Verlag Heidelberg.
Ejermo, O. (2005). Technological diversity and Jacob's externality hypothesis revisited. *Growth and Change*, 36 (2), 167-195.
Fleming, L. & Sorenson, O. (2001). Technology as a complex adaptive system: evidence from patent data. *Research Policy*, 30, 117-132.
Galindo-Rueda, F. & Verger, F. (2016). OECD Taxonomy of Economic Activities Based on R&D Intensity. *OECD Science, Technology and Industry Working Papers*, 2016/04. Paris: OECD.
Giannitsis, T. & Kager, M. (2009). *Technology and Specialization: Dilemmas, Options, Risks?* Brussels: European Commission.
Glaeser, E., Kallal, H., Scheinkman, J. & Shleifer, A. (1992). Growth in Cities. *The Journal of Political Economy*, 100 (6), 1126-1152.
Grilliches, Z. (1990). Patent statistics as economic indicators: a survey. *Journal of Economic Literature*, 28, 1661-1707.
Ha, S.H., Liu, W., Cho, H. & Kim, S.H. (2015). Technological advances in the fuel cell vehicle: Patent portfolio management. *Technological Forecasting and Social Change*, 100, 277-289.
Jacobs, J. (1969). *The Economy of Cities*. NY: Random House.
Jacobs, J. (1984). *Cities and the Wealth of Nations: Principles of Economic Life*. NY: Random House.
Jaffe, A., Trajtenberg, M. & Henderson, R. (1993). Geographical localization of knowledge spillovers as evidenced by patent citations. *The Quarterly Journal of Economics*, 108 (3), 577-598.
Khramova, E., Meissner, D. & Sagieva, G. (2013). Statistical patent analysis indicators as a means of determining country technological specialization. *Working Papers Series: Science, Technology and Innovation, WP BRP 09/STI/2013*. Retrieved November 10, 2018 from: https://www.hse.ru/data/2013/04/10/1297571825/09STI2013.pdf.
Kogler, D.F., Heimeriks, G. & Leydesdorf, L. (2018). Patent portfolio analysis of cities: statistics and maps of technological inventiveness. *European Planning Studies*, 26 (11), 2256-2278.
Kogler, D.F., Rigby, D. & Tucker, I. (2013). Mapping Knowledge Space and Technological Relatedness in US Cities. *European Planning Studies*, 21 (9), 1374-1391.
O'hUallachain, B. (1999). Patent places: size matters. *Journal of Regional Science*, 39 (4), 613-636.
Pattel, P. & Pavitt, K. (1997). The technological competencies in the world's largest firms: complex and path dependent, but not too much variety. *Research Policy*, 26, 141-156.
Pianta, M. & Meliciani, V. (1996). Technological specialization and economic performance in OECD countries. *Technology Analysis & Strategic Management*, 8 (2), 157-174.
Rigby, D. (2015). Technological relatedness and knowledge space: entry and exit of US cities from patent classes. *Regional Studies*, 29 (11), 1922 – 1937.

Schmoch, U. (2008). Concept of a technology classification for country comparisons. *Final Report to the World Intellectual Property Organization (WIPO)*, Fraunhofer Institute for Systems and Innovation Research, Karlsruhe.

Strumsky, D., Lobo, J. & van der Leeuw, S. (2012). Using patent technology codes to study technological change. *Economics of Innovation and New Technology*, 21 (3), 267-286.

Van Looy, B., Vereyen, C. & Schmoch, U. (2014). *Patent Statistics: Concordance IPC V8 – NACE Rev.2*. Eurostat.

Vlckova, J., Kasprikova, N. & Vlckova, M. (2018). Technological relatedness, knowledge space and smart specialization: the case of Germany. *Moravian Geographical Reports*, 26 (2), 95-108.

Xia, L. & Hu, J.-Y. (2014). Analysis of the relationship between specialization and innovation of technology in Chinese cities. *Applied Mechanics and Materials*, 522 – 524, 1577-1580.

Gokhberg, L.M. (2003). *Statistika nauki (Statistics of Science)*. Moscow: TEIS (in Russian).

Rospatent (2018). *Annual report for 2017* (in Russian).

Categorization model of Spanish scientific journals in social sciences and humanities

Daniela De Filippo[1]; Rafael Aleixandre-Benavent[2] and Elías Sanz-Casado[3]

[1] *dfilippo@bib.uc3m.es*
INAECU Institute (Research Institute for Higher Education and Science), Carlos III University of Madrid, Calle Madrid 126, Getafe, 28903, Madrid (Spain)

[2]*Rafael.Aleixandre@uv.es*
Ingenio Institute UISYS, Joint Research Unit, Universitat de València-CSIC, Plaza Cisneros 4 46003, Valencia (Spain)

[3]*elias@bib.uc3m.es*
INAECU Institute (Research Institute for Higher Education and Science), Dept. of Library and Informatio Science, Carlos III University of Madrid, Calle Madrid 126, Getafe, 28903, Madrid (Spain)

Abstract

Social sciences and humanities (SSH) present differences from other scientific areas due to the nature of their research (local or national), the documentary typology of its scientific writing (mainly books and monographs), its habits of collaboration (individual or with little institutional collaboration) and language (preferably vernacular). These particularities make the production visibility in SSH smaller than that in other areas. Likewise, this limits the evaluation of journals, researchers and institutions because it is not possible to obtain indicators to determine and compare productivity, collaboration or impact; it is also difficult to build classifications of scientific journals according to their quality that would make it possible to draw comparisons between journals. To overcome these limitations, we present a classification of national scientific journals with 2 dimensions (impact and visibility) based on quantitative criteria. To obtain indicators in each of these dimensions, databases such as Science Citation Index (SCI), Social Science Citation Index (SSCI), Arts & Humanities Citation Index (AHCI) and Emerging Sources Citation Index from Web of Science Core Collection, Scopus, SciELO, Google Metrics and Information Matrix for the Analysis of Journals (MIAR) are used as sources of information. From these sources, we obtain indicators of coverage, citations, h-index and quartile. The Spanish Foundation for Science and Technology (FECYT) uses this classification to rank the journals that have obtained the "Quality Seal" and its implementation is being considered by the Spanish evaluation agencies.

Introduction

The publication of the results of research in the social sciences and humanities (SSH) presents differences from publications in other areas. The research should be local or national, so results from SSH studies obtained in one country may not always be very useful to researchers in other countries (Nederhof, 2006); the documentary typology is different because chapters in books and monographs are more frequently used as publication channels and are cited more often than journal articles (Hicks, 2004; Nederhof, 2006); habits of collaboration are different (individual or with little institutional collaboration); research is often written in the vernacular and is seldom covered in the bibliometric databases (Chi, 2014) or even published in other publication channels that are not covered at all (for example, reports and other publications directed to regional readership); SSH researchers write not only for scholarly readers but also for the lay public (Hicks, 2004); and this type of literature is usually not included in the databases used for bibliometric analyses. Additionally, citation behaviour is different in the SSH disciplines and, for example, the age of references is remarkably high, so it may not be appropriate to apply the citation windows of two years considered for the calculation of impact factor (Glänzel, 1999).
In SSH, many national scientific publications, unlike publications from other areas of science and technology, are not indexed in the databases that are usually used by the evaluation

committees. Therefore, the evaluation agencies do not usually have specific mechanisms to categorize and order national scientific journals (Moed, 2004). Bibliometric indicators of publication output and citation impact have been widely used in the evaluation of scientific journals because such indicators offer high transparency and strong legitimacy by allowing the researchers themselves to scrutinize the recounts of published papers, their citations and, therefore, the calculation of indicators (Ahlgren, Colliander & Persson, 2012). In addition, there are several sources from which we can calculate these indicators, such as WoS, Scopus and SciELO (Scientific Electronic Library Online), as well as other complementary derived sources, such as Journal Citation Reports or Scimago Journals and Country Ranks.

When examining the international experiences of the classification of journals in SSH, one of the first classifications to be mentioned is the European Reference Index for Humanities (ERIH), an initiative launched by the European Science Foundation (ESF) in 2001 and published in 2007. In France, the Agence d'Evaluation de la Recherche et de l'Enseignement Superieur (AERES) published a list in 2008 with 6,305 journals organized in three classes (A, B, and C). In Australia, in 2010, the government launched the Excellence in Research for Australia exercise, an expert-based classification of journals including more than 20,000 titles. Other countries such as Brazil and other Southern American countries, Italy (Ferrara and Bonacorsi, 2017), Taiwan, The Netherlands, Norway, Denmark and Sweden have also made schemes to classify and rank journals in SSH (Ahlgren, Colliander and Persson 2012; Ingwersen and Larsen 2014; Hammarfelt and De Rijcke 2015)

In Spain, the RESH is a journal evaluation system with more than 2,000 SSH journals classified in four classes (A, B, C, and D) in addition to an excellence class (Gimenez-Toledo, Roman-Roman and Alcain-Partearroyo, 2007). Another classification system, CIRC (Clasificación Integrada de Revistas Científicas), categorizes more than 20,000 journals in four categories (A+, A, B, C) (Torres-Salinas et al. 2010).

Certainly all of these systems have made an important contribution to the criteria for evaluating SSH journals. However, they also have some limitations. In some cases, the lack of continuity in the collection of information prevents us from having updated information, in other cases the classification allows us to assign journals to a certain group but not to carry out rankings within each group. On the other hand, although the use of bibliometric indicators has been highly criticised, numerous studies show that there is a correlation with the results obtained through consultation with experts.

Considering these limitations, our proposal tries to advance towards the elaboration of a methodology that allows to rank journals that have already passed a quality threshold (measured quantitatively and qualitatively).

Evaluation of Spanish journals in social science and humanities, the FECYT Quality Seal

Since 2006, the Spanish Foundation for Science and Technology (FECYT) has been carrying out the ARCE project (FECYT, 2018) with the aim of contributing to the professionalisation and internationalisation of Spanish scientific journals. These journals undergo an evaluation process in accordance with international standards and those that surpass it are awarded the Quality Seal. So far, 298 journals from different areas of knowledge have obtained this Seal, with a greater representation of those belonging to the Social Sciences and Humanities.

This Seal, backed by more than 10 years of FECYT experience, is a first-rate distinction in Spain and is also supported by national and international experts who have formed part of the evaluation committees of the calls made.

The editorial and scientific quality indicators required, in the case of some journals, have produced a real change of strategy in their editorial processes. These quality indicators are applied according to the areas of knowledge of each journal.

Once the accreditation process was consolidated, a new challenge arose: to provide evaluation agencies with a list of the highest quality Spanish journals in order to incorporate them as a merit in the curricular evaluation of researchers.
To this end, FECYT has convened a committee of experts to generate a methodology to categorise journals already recognised with the Quality Seal and to offer a list (ordered according to merit) in each scientific category, especially in Social Sciences and Humanities.
The authors of this paper have been working along these lines since 2015. After several rounds of methodological discussion and with the validation of different external experts, the application of the methodology detailed in the following sections has been agreed for the classification of Spanish journals. The main objective of this paper is to examine the robustness and applications of a classification system of Spanish SSH journals based on journal-level indicators. This work takes advantage of and integrates existing evaluation resources that have been previously evaluated positively by the evaluation agencies to build a classification of scientific journals in SSH.

Methods

For the evaluation and classification of accredited journals under the FECYT's Seal of Quality, we propose a methodology based on two dimensions: the analysis of journals' impacts and the analysis of their visibility.
To obtain indicators in each of these dimensions, databases such as Science Citation Index (SCI), Social Science Citation Index (SSCI), Arts & Humanities Citation Index (AHCI) and Emerging Sources Citation Index (ESCI) from Web of Science Core Collection (WoS), Journal Citation Reports, Scopus, SciELO, Google Metrics, and Information Matrix for the Analysis of Journals (MIAR) are used as sources of information. From these sources, we obtain indicators of coverage in each database, citations, H index and quartile.
It should be noted that only the journals that already had the FECYT Quality Seal have been considered (298 Spanish journals). This is because it has been considered fundamental to start with a selection of journals that meet basic quality criteria. In particular, this work has been carried out with the journals of social sciences (127) and humanities (110).

Proposed model for the categorization of journals: dimensions and indicators

A very important decision in dating counts is the selection of the "citation window", that is, the number of years that should be considered to quantify citations. Some studies have concluded that a period of 2 years, used in the calculation of the impact factor, may be enough in areas such as biomedical studies (Campanario, 2011), physics and some life sciences (Adams, 2005). However, in SSH and mathematics, in which the citation dynamics are slower and therefore need more time to receive citations, this period is too short, and it is necessary to extend it so that the majority of the publications can be recognized and cited (Vanclay; 2008, Campanario, 2011; Waltman and Van Eck, 2012; Dorta-González and Dorta-González, 2013). Along these lines, the Journal Citation Reports of Web of Science introduced a new indicator in 2007 for the journals included in its coverage, namely, the impact factor of 5 years, which includes a citation window of 5 years and serves to complement the impact factor in the short term of 2 years (Jacso, 2009). Given the variability in the citation windows according to the thematic area, this study has considered a citation window of 5 years, since it is better suited to citations in SSH and does not harm others such as biomedical studies in which 2 or 3 years would be enough.
The indicators obtained from the different sources considered in this methodology have been grouped into two dimensions according to their characteristics: Impact and Visibility.

Both dimensions make up the final score of each journal that is composed of the percentage reached in each indicator. In order to fix the percentage values, several previous exercises have been carried out that were validated by experts (Sanz-Casado et al., 2017).

Dimension 1: Impact

The first dimension includes the indicators directly linked to the citations. **This dimension represents 80%** of the score obtained by the journal and contains three groups of indicators: citations, H indexes and quartiles.

1.1. Citations

- **Citations in SCI, SSCI & AHCI**. To collect the number of citations that the selected Spanish journals have received from publications indexed in these databases, the tool "cited reference search" has been used. The name and variants of the journal have been searched in the field "cited work". The search was limited to the last 5 years in "cited year(s)".
- **Citations in Scopus**. To obtain the citations of a journal in Scopus, we access the "advanced search" function and in the search box enter the name of the journal in parentheses preceded by the tag REFSRCTITLE. The search period is limited to the last 5 years the using the REFPUBYEAR tag.
- **Citations in SciELO**. The SciELO initiative (Scientific Electronic Library Online) was created in response to the lack of indexation of a large part of scientific journals in Latin America and the need for these journals to be included in quality bibliographic systems and for access to full-text articles. Currently, this database is the only instrument for measuring the impact of such journals because most of them are not indexed in other important databases such as WoS Core Collection and Scopus (Alfonso, Rodríguez-Morales and Mayta-Tristán, 2009). The search methodology is similar to the search for citations in SCI, SSCI & AHCI, but here, the first selection is the SciELO Citation Index. The search was also limited to the last 5 years.
- **Citations in Emerging Sources Citation Index (ESCI)**. ESCI is a database composed of journals that are being evaluated for their next incorporation into the main collections of the Web of Science Core Collection. The search method is like that of SCI, SSCI & AHCI, but it limits the search to ESCI database.

The citations obtained from the different sources are added together, and the value obtained represents 60% of the total score that the journal will obtain.

1.2. h-Index

- **h-index in WoS**. The h-index is considered a robust indicator for determining the impact of scientific publications, researchers, etc., because both the works not cited and those highly cited do not significantly affect the value of the indicator (Braun, Glänzel and Schubert, 2006). This indicator tends to assess a prolonged scientific effort along a trajectory (Braun, Glänzel and Schubert, 2006). Furthermore, one of its strengths is that it combines the effect of quantity (number of published articles) with quality (number of citations received) (Norris and Oppenheim, 2010). The h-index of each journal has been obtained by searching the WoS (SO="name of the journal") and has been limited to the last 5 years.
- **h-index in SJR**. The h-index in Scimago Journal & Country Rank is offered entering the name or ISSN of the journal in the search engine of the application
- **h5-Index in Google Scholar Metrics**. This indicator is obtained consulting the Google Scholar Metrics website

The sum of the values obtained by the h-index represents 10% of the journal's score.

1.3. Quartiles

- **Quartile in JCR**. The quartile of a journal is considered an indicator of quality and allows us to know which of the publications is best valued within each category in which the WoS classifies the journals. Since a journal may be included in more than one WoS category, it has been considered in the highest quartile achieved. If the journal is in the first quartile (Q1), where it has the maximum visibility, the score obtained is 100 points. If its location is in the second quartile (Q2), the score obtained by the journal is 75. If it is in the third quartile (Q3) the score is 50, and finally, in the fourth quartile, the journal score is 25 points.
- **Quartile in SJR**. The database of Scimago Journal & Country Rank (SJR) offers the bibliometric information of more than 17,000 academic and professional journals based on data extracted from Elsevier's Scopus database. As in JCR, one of these measures is the quartile occupied by journals in their respective thematic areas (Jacsó, 2010; Mañana Rodríguez, 2014). The quartile in Scimago Journal & Country Rank is offered after entering the name or ISSN of the journal in the search engine of the application (http://www.scimagojr.com/). As in the previous case, if the journal is located in the first quartile (Q1) the score obtained is 100. If its location is in the second quartile (Q2), the score of the journals is 75, and 50 if it is in the third quartile (Q3); finally, in the fourth quartile, the score obtained is 25.

The sum of the values obtained by the h-index represents 10% of the journal's score.

Dimension 2: Visibility

The second dimension is visibility, understood as the presence of the journal in different databases of international prestige. The following indicator has been considered for this purpose.

- **Secondary Composite Index Broadcasting** (ICDS) of MIAR (Information Matrix for the Analysis of Journals). MIAR is a database created at the University of Barcelona (http://miar.ub.edu/about-miar), with the purpose of determining the visibility of scientific journals based on their presence in scientific databases, both national and international, as well as in multidisciplinary repertoires. MIAR groups journals in large scientific areas that are in turn subdivided into more specific fields. The visibility of the journals is quantified through the ICDS, an indicator that shows the visibility of the journal in different scientific databases of international scope, or their failure in repertoires' evaluation of periodicals. More than 100 databases (as WoS, Scopus, MEDLINE, ERIH plus, etc.) are consulted. A high ICDS means that the journal is present in different sources of information of international relevance. For the calculation of ICDS, different criteria are established (http://miar.ub.edu/about-icds).

As mentioned in the methodology of the database, since 2016, changes have been made in the calculation of the indicator. Now, the value is rounded to the first decimal. Regarding the score given for the presence of journals in bibliographical databases, there are two changes: 3+2 points are given when a journal is covered by two or more abstracting and indexing databases without distinction between specialized or multidisciplinary and an additional +1 point granted to journals that are indexed simultaneously in Scopus and in classic WoS indexes (Arts and Humanities Citation Index, Science Citation Index Expanded and Social Sciences Citation Index) (MIAR, 2018).

The sum of the values obtained by the ICDS of MIAR represents 20% of the journal's score.

Calculation of the indicator for the categorization of Spanish journals

The first step in calculating the final score of each journal, and therefore its position within an sorted list, is to group all of them by subject area. The indicators are then obtained according to the following criteria:

A) Citation Indexes

- Retrieval of the number of citations of each journal from each database.
- Sum of all citations received by each journal.
- Descending order of the journals of each thematic category according to the total number of citations received.
- Rescaling of the values of citations obtained from each journal according to the maximum weight of the dimension (60%), in such a way that the most cited journal is assigned 60 points, and the values for the rest of journals are calculated according to the maximum value obtained.

B) H-index

- Obtaining the H-index values of each magazine from the 3 databases (WoS, SJR and Google Scholar Metrics).
- The H-index value of the journal in each database is divided by the highest value of the journal in that database and the result is multiplied by 1/3 of the total weight of that indicator (10%).
- The final H-index score of each journal is the result of the sum of the H-index values that that journal has obtained in each database.

C) Quartiles

- Obtaining the position of the journals in the quartiles of each database considered.
- Assignment of normalized values to each quartile: Q1=100, Q2=75, Q3=50, and Q4=25.
- The normalized quartile value of the journal in each database is divided by the highest value of the journal in that database and the result is multiplied by 1/2 of the total weight of that indicator (10%).
- The final score of each journal is the result of the sum of the values that the journal has obtained in each database.

D) Visibility

- Obtaining ICDS values from MIAR for each journal (http://miar.ub.edu/)
- Sorting journals in descending order according to the value of the ICDS of each one.
- Rescaling of the ICDS values obtained from each journal according to the maximum weight of the visibility dimension (20%): The journal with the highest ICDS will have 20 points, and the rest of the journals will be assigned points with respect to that maximum value.

Final Score

The final value obtained by each journal will be the equivalent to the values of each of the previous phases: citations, H-index, quartile, ICDS. The maximum to reach is always 100.

Quartile distribution

Finally, once the final scores of the journals in each field have been calculated, they have been organized by quartiles, starting from a homogeneous division of each field into four equal parts taking into account the total number of journals and the final score obtained by each one of them. The result of each quartile follows (table 3):

- Quartile A, contains 25% of the journals with the highest score.
- Quartile B, contains 25% of the journals with a high average score.
- Quartile C, contains 25% of the journals with low average score.

• Quartile D, contains 25% of the journals with the lowest score.

Results

The model was checked in 237 Spanish journals with the FECYT Quality Seal from the following fields: Social, political, behavioural and educational sciences; economic and business sciences; law; history, geography and arts; philosophy, philology and linguistics.

In table 1, an example has been made with the journals included in the field *Education*, with 38 journals. The table shows the absolute values of each dimension and the determined weighting for the first 10 journals. It is appreciated that the journal *Revista de Educación* reaches the highest values in all dimensions since it is a journal included in WoS and Scopus, with the high number of citations coming mainly from other journal indexed in WoS.

Table 1. Citation indicators from Education journals

Journal Title	*Citation SCI, SSCI&AHCI*	*Citation SCOPUS*	*Citation ESCI*	*Citation SCIELO*	*Total citation*	*Total (60%)*
Revista de Educación	539	1.755	148	67	2509	**60,00**
Revista de Psicodidáctica	515	241	282	56	1094	**26,16**
Enseñanza de las Ciencias	192	677	134	39	1042	**24,92**
Revista Iberoamericana de Educación (versión monográfica)	57	677	103	66	903	**21,59**
RIE. Revista de Investigación Educativa	147	465	237	31	880	**21,04**
RETOS. Nuevas Tendencias en Educación Física, Deporte y Recreación	231	279	304	30	844	**20,18**
Cultura y Educación	246	336	147	28	757	**18,10**
Revista Eureka Sobre Enseñanza y Divulgación de las Ciencias	121	207	161	20	509	**12,17**
Profesorado. Revista de Currículum y Formación del Profesorado	69	216	121	21	427	**10,21**
Revista Interuniversitaria de Formación del Profesorado	34	291	52	34	411	**9,83**

Table 2 shows the H index values obtained for the *Education* journals and the calculation of their weighting. The first 10 journals of the discipline are shown. The *Revista de Psicodidáctica* stands out with high values of H-Index in WoS and the *Revista de Educación* moves to the second position.

Table 2. H-index indicators from Education journals

Journal Title	*H-Index WoS*	*Calculation H-Index WOS*	*H-Index SJR*	*Calculation l H-Index SJR*	*Google S.Metrics (h5-Index)*	*Calculation h5-Index GSM*	*TOTAL H-INDEX (10%)*
Revista de Psicodidáctica	13	3,33	12	3,33	18	2,40	9,07
Revista de Educación	6	1,54	10	2,78	25	3,33	7,65
Educación XX1	5	1,28	6	1,67	14	1,87	4,82

RU&SC. Revista de Universidad y Sociedad del Conocimiento	0	0,00	7	1,94	19	2,53	4,48
Cultura y Educación	6	1,54	9	2,50	0	0,00	4,04
Porta Linguarum	4	1,03	5	1,39	9	1,20	3,61
Enseñanza de las Ciencias	0	0,00	6	1,67	14	1,87	3,53
RICYDE. International Journal of Sport Science	0	0,00	6	1,67	13	1,73	3,40
RIE. Revista de Investigación Educativa	0	0,00	5	1,39	15	2,00	3,39
Relieve. Revista Electrónica de Investigación y Evaluación Educativa	0	0,00	7	1,94	10	1,33	3,28

When considering the quartiles, it is also the journal *Revista de Psicodidáctica* that leads the rankings for being in Q1 in SJR and Q2 in JCR (table 3).

Table 3. Quartile indicators from Education journals

Journal Title	*Quartile JCR*	*Calculation Q JCR*	*Quartile SJR*	*Calculation Q SJR*	*TOTAL CUARTIL (10%)*
Revista de Psicodidáctica	75	5	100	5	**10,00**
Cultura y Educación	25	1,67	100	5	**6,67**
Revista de Educación	50	3,33	50	2,5	**5,83**
Porta Linguarum	25	1,67	75	3,75	**5,42**
Educación XX1	25	1,67	50	2,5	**4,17**
RIE. Revista de Investigación Educativa	0	0,00	50	2,5	**2,50**
Enseñanza de las Ciencias	0	0,00	50	2,5	**2,50**
RU&SC. Revista de Universidad y Sociedad del Conocimiento	0	0,00	50	2,5	**2,50**
RICYDE. Revista Internacional de Ciencias del Deporte = International Journal of Sport Science	0	0,00	50	2,5	**2,50**
Revista Eureka Sobre Enseñanza y Divulgación de las Ciencias	0	0,00	50	2,5	**2,50**

Regarding visibility, given a maximum of 11 points, the journal *Revista de Educación* reaches the maximum do to, in addition to its quality, it has a long history (51 years old) (table 4)

Table 4. Visibility indicators from Education journals

Journal Title	*MIAR*	*TOTAL MIAR (20%)*
Revista de Educación	11	**20,00**

RIE. Revista de Investigación Educativa	10	**18,18**
Teoría de la Educación. Revista Interuniversitaria	10	**18,18**
Apunts. Educación Física y Deportes	10	**18,18**
Revista Interuniversitaria de Formación del Profesorado	10	**18,18**
Revista Electrónica Interuniversitaria de Formación del Profesorado	10	**18,18**
Bordón. Revista de Pedagogía	10	**18,18**
Pedagogía Social. Revista Interuniversitaria	10	**18,18**
Educatio Siglo XXI	10	**18,18**
Revista Complutense de Educación	9,9	**18,00**

Once the indicators of each dimension have been calculated, we proceed to the total count and the ordering of journals according to the values obtained. Table 5 shows the first 10 *Education* journals and is observed that *Revista de Educación* reaches the first position with 93.48 points.

Table 5. Final SCORE from *Education* journals

Journal Title	***Total Score***
Revista de Educación	**93,48**
Revista de Psicodidáctica	**61,23**
Cultura y Educación	**48,63**
Enseñanza de las Ciencias	**47,32**
RIE. Revista de Investigación Educativa	**45,11**
Revista Iberoamericana de Educación (versión monográfica)	**41,33**
RETOS. Nuevas Tendencias en Educación Física, Deporte y Recreación	**39,77**
Educación XX1	**37,90**
Revista Eureka Sobre Enseñanza y Divulgación de las Ciencias	**34,68**
RU&SC. Revista de Universidad y Sociedad del Conocimiento / International Journal of Educational Technology in Higher Education	**32,25**

Finally, once the final scores of the journals in each field have been calculated, they have been organized by quartiles, starting from a homogeneous division of each field into four equal parts considering the total number of journals and the final score obtained by each one of them.

Final Considerations

This work integrates indicators of impact and visibility to obtain a model of the classification and categorization of journals on SSH based on journal-level indicators. The sources and indicators used to build this model have the advantage that have been positively tested by both evaluation agencies and researchers in scientific evaluation. Furthermore, we start from a collection of journals previously accredited by the Spanish evaluation agency for Science and Technology FECYT.

In this line, unlike other methods used for the evaluation of SSH journals in Spain, the methodology used has the following advantages:

- analysis of the total number of journals that have already obtained the "quality seal" de FECYT.
- annual update
- inclusion of numerous sources of information
- large citation window
- collection of citations from maistream journals, even if the journal itself is not indexed in an international database
- classification of journals in quartiles

- positioning within each quartile
- inclusion of a visibility index based on presence in numerous databases

Journal rankings based on bibliometric indicators are generally seen as inadequate. The limitations of existing bibliometric databases in the case of SSH have been carefully discussed in the literature (Nederhof and Zwaan 1991; Nederhof and Noyons 1992; Archambault et al. 2006; Nederhof 2006; Hicks and Wang 2009; Hellqvist 2010; Linmans 2010), and the very use of citations as the basis for journal ranking in SSH, whatever the specific metrics and the database adopted, has been the object of severe criticism (Moed, Luwel and Nederhof 2002; Campbell, Goodacre and Little 2006; Jarwal, Brion and King 2009). One problem in the process of ranking journals based on citations is that of judging the statistical significance of the difference between two scores. Nevertheless, Elkins et al. (2010) calculate the pairwise correlations between the ISI journal impact factor and three other journal citation indexes, and the correlations of four indexes were found to range from strong to very strong, providing evidence of convergent validity, that is, closely related average journal citations per article. From a purely statistical perspective, it does not seem to matter which index might be used to capture the impact of citations, despite substantial differences in constructing the different measures of citations. Another problem is that the numerical difference between the impact factors of two journals can be so small that it is unlikely to be statistically significant, and this fact can affect their order in the rankings of journals (Moosa, 2016). Vasen and Lujano Vilchis (2010) compared the guidelines for classifying journals in the social sciences implemented in three Latin America countries. They concluded that the tendencies show an evolution of the evaluation system of journals towards a model based on citation indexes and a relative reduction of the importance of other databases linked to specific fields of knowledge or to particular regions (Vasen & Lujano Vilchis, 2017).

Despite these drawbacks, management and decision making in SSH have benefited from the role played by bibliometric analysis studies, and several works have been published employing bibliometric indicators to evaluate journals and to build journal's classifications (Gimenez-Toledo, Roman-Roman and Alcain-Partearroyo, 2007; Ferrara and Bonacorsi, 2017; Ahlgren, Colliander and Persson 2012; Ingwersen and Larsen 2014; Hammarfelt and De Rijcke 2015).

Logically, this methodological proposal is not exempt from limitations. The application of bibliometric methods for classifying and ranking journals in SSH is problematic and yielded unsatisfying results, so even bibliometricians warn against applying bibliometric methods to these areas (Nederhof and Zwaan, 1991; Glänzel and Schoepflin, 1999; Ochsner, Hug and Galleron, 2017). First, it is important to keep in mind that the number of citations is an approximate estimation of the scientific relevance of a journal, but currently it is commonly an accepted indicator or indirect method to measure the quality of research. Second, international databases used in this work have a coverage bias that harmed national or regional journals. However, the practice of integrating several different sources allows this limitation to be partially overcome.

In conclusion, the classification systems of publications are of great importance because they guide researchers to the journals in which it is desirable to publish. The publication patterns in SSH are different from other scientific areas, so journals must be evaluated and ranked with a rigorous methodology that considers a high number of criteria from accredited sources. The application of bibliometric methods to the evaluation and ranking of scientific journals in SSH has been found to be a valid approach. The integration of several measures and indicators of impact and visibility lends accuracy and robustness to the evaluation. In addition, in counting citations, we have considered a citation window of 5 years to be more appropriate for SSH. The proposed classification held great interest for evaluation agencies endeavouring to know the visibility and impact of Spanish journals in SSH, and this classification system could be applied in other countries using similar or alternative sources.

Acknowledgements
This study was sponsored by the Spanish Foundation for Science and Technology (FECYT), which carries out a permanent evaluation of the journals and awards the Quality Seal.

References

Adams, J. (2005). Early citation counts correlate with accumulated impact. *Scientometrics*, 63(3), 567-581.

Ahlgren, P., Colliander, C. & Persson, O. (2012). Field normalized citation rates, field normalized journal impact and Norwegian weights for allocation of university research funds. *Scientometrics*, 92, 767-780.

Alfonso, J.; Rodriguez-Morales, A.J. & Mayta-Tristán, P. (2009). Preliminary bibliometric evaluation of scientific publications produced in Latin America in the field of tropical and infectious diseases using SciELO. *Journal of Infection in Developing Countries*, 3(3), 247-249.

Archambault, E., Vignola-Gagne, E., Cote, G., Larivie´re, V. & Gingras Y. (2006) Benchmarking scientific output in the social sciences and humanities: the limits of existing databases. *Scientometrics*, 68(3), 329–42.

Braun, T.; Glänzel, W. & Schubert, A. (2006). A Hirsch-type index for journals. *Scientometrics*, 69(1), 169–173

Campanario, J. M. (2011). Empirical study of journal impact factors obtained using the classical two-year citation window versus a five-year citation window. *Scientometrics*, 87, 189–204.

Campbell, K., Goodacre, A. & Little, G. (2006). Ranking of United Kingdom law journals: an analysis of the research assessment exercise 2001 submissions and results. *Journal of Law and Society*, 33, 335–363.

Chi, P.S. (2014). Which role do non-source items play in the social sciences? A case study in political science in Germany. *Scientometrics*, 101(2), 1195–1213.

Dorta-González, P. & Dorta-González. MI. (2013). Comparing journals from different fields of science and social science through a JCR subject categories normalized impact factor. *Scientometrics*, 95, 645-672.

Elkins, M. R., Maher, C. G., Herbert, R. D., Moseley, A. M., & Sherrington, C. (2010). Correlation between the journal impact factor and three other journal citation indices. *Scientometrics*, 85, 81–93.

Ferrara, A. & Bonaccorsi, A. (2016). How robust is journal rating in Humanities and Social Sciences? Evidence from a large-scale, multi-method exercise. *Research Evaluation*, 25, 279-291.

FECYT (2018) *Proyecto ARCE*. Accesible en: https://www.fecyt.es/es/recurso/arce

Giménez-Toledo, E., Román-Román, A. & Alcaín-Partearroyo, D. (2007). From experimentation to coordination in the evaluation of Spanish scientific journals in the humanities and social sciences. *Research Evaluation*, 16 (2), 137–48.

Glänzel, W. & Schoepflin, U. (1999). A bibliometric study of reference literature in the sciences and social sciences". *Information Processing and Management*, 35, 31-44.

Hammarfelt, B. and De Rijcke, S. (2015). Accountability in context: effects of research evaluation systems on publication practices, disciplinary norms, and individual working routines in the Faculty of Arts at Uppsala University. *Research Evaluation*, 24(1), 63–77.

Hellqvist, B. (2010). Referencing in the humanities and its implications for citation analysis. *Journal of the American Society for Information Science and Technology*, 61(2), 310–318.

Hicks, D. (2004). The four literatures of social science. In: Moed H. (Ed). *Handbook of quantitative science and technology research*. Dordrecht: Kluwer Academic, pp. 473-496.

Hicks, D. & Wang, J. (2011). Coverage and overlap of the new social science and humanities journal lists. *Journal of the American Society for Information Science and Technology*, 62(2), 284–294.

Ingwersen, P. & Larsen, B. (2014). Influence of a performance indicator on Danish research production and citation impact 2000–12. *Scientometrics*, 101(2), 1325-1344.

Jacsó. P. (2010). Comparison of journal impact rankings in the Scimago Journal & Country Rank and the Journal Citation Reports databases. *Online Information Review*, 34(4), 642-657.

Jarwal, S. D., Brion, A. M. & King, M. L. (2009). Measuring research quality using the Journal Impact Factor, citations and ranked journals. Blunt instruments or inspired metrics? *Journal of Higher Education Policy and Management*, 31(4), 289–300.

Linmans, A. J. M. (2010). Why with bibliometrics the Humanities does not need to be the weakest link indicators for research evaluation based on citations, library holdings, and productivity measures, *Scientometrics*, 83, 337–54.

Mañana Rodriguez J. (2014). A critical review of Scimago Journal & Country Rank. *Research Evaluation*, 24(4), 343-354.

MIAR. Matriz de Información para el Análisis de Revistas. (2018). *Sobre el ICDS*. Accesed in: http://miar.ub.edu/about-icds.

Moed H. (2004). *Handbook of quantitative science and technology research*. Dordrecht: Kluwer Academic.

Moed, H. F., Luwel, M. & Nederhof, A. J. (2002). Towards research performance in the humanities. *Library Trends*, 50, 498–452.

Moosa, I. A. (2016). A Critique of the Bucket Classification of Journals: The ABDC List as an example. *Economic Record*, 92, 448–463.

Nederhof, A.J. (2006). Bibliometric monitoring of research performance in the social sciences and the humanities: a review. *Scientometrics*, 66, 81-100.

Nederhof, A. J. & Noyons, E. C. M. (1992). International Comparison of Departments' Research Performance in the Humanities. *Journal of the American Society for Information Science*, 43(3), 249–56.

Nederhof, A. J. & Zwaan, R. A. (1991). Quality judgments of journals as indicators of research performance in the humanities and the social and behavioral sciences. *Journal of the American Society for Information Science,* 42(5), 332-340.

Norris. M. & Oppenheim, C. (2010). The h-index: a broad review of a new bibliometric indicator. *Journal of Documentation*; 66(5), 681-705.

Ochsner, M.; Hug, S & Galleron, I. (2017). The future of research assessment in the humanities: bottom-up assessment procedures. *Palgrave Communications* 3, Article number: 17020. doi:10.1057/palcomms.2017.20.

Sanz-Casado, E.; De Filippo, D. & Aleixandre-Benavent, R. (2017). *Guía metodológica de clasificación de revistas de Ciencias Sociales y Humanidades.* FECYT, Madrid Avalaible at: https://www.fecyt.es/es/noticia/guia-metodologica-de-clasificacion-de-revistas-de-ciencias-sociales-y-humanidades

Rodríguez-Gairín, J. M.; Somoza-Fernández, M. & Urbano, C. (2011). MIAR: hacia un entorno colaborativo de editores, autores y evaluadores de revistas. *El profesional de la información*. 20(5), 589-595.

Torres-Salinas, D., Bordons, M., Giménez-Toledo, E., Delgado-López-Cózar, E., Jiménez-Contreras, E. & Sanz-Casado, E. (2010). Clasificación integrada de revistas científicas (CIRC): propuesta de categorización de las revistas en ciencias sociales y humanas. *El Profesional de la Información,* 19(6), 675-684.

Vanclay, J. K. (2012). Impact factor: Outdated artefact or stepping-stone to journal certification? *Scientometrics*, 92(2), 211-238.

Vasen, F. & Lujano Vilchis, I. (2017). Sistemas nacionales de clasificación de revistas científicas en América Latina: tendencias recientes e implicaciones para la evaluación académica en ciencias sociales. *Revista Mexicana de Ciencias Políticas y Sociales*, 62, 199–228.

Waltman, L., & Van Eck, N. J. (2012). A new methodology for constructing a publication-level classification system of science. *Journal of the American Society for Information Science and Technology*, 63(12), 2378-2392.

Article Level Classification of Publications in Sociology: An Experimental Assessment of Supervised Machine Learning Approaches

Joshua Eykens[1], Raf Guns[1], and Tim C. E. Engels[1]

[1]joshua.eykens@uantwerpen.be
[1]raf.guns@uantwerpen.be
[1]tim.engels @uantwerpen.be

[1]Centre for R&D Monitoring (ECOOM), Faculty of Social Sciences, University of Antwerp, Middelheimlaan 1, B-2020 Antwerp (Belgium)

Abstract

The purpose of this experiment is to assess whether and to what extent it is feasible to make use of supervised machine learning to classify social science journal articles into fine-grained disciplinary categories. Classifying scientific articles according to disciplines is most commonly done by making use of a proxy such as Clarivate Analytics' Web of Science journal level Subject Categories. Past research has shown that this approach does not come without limitations. Classifications based on textual data might be more appropriate in this case. In this paper we make such an attempt using titles and abstracts. We test four different supervised machine learning algorithms and assess their accuracy when it comes to granularly classifying sociology publications based on textual information. Our results show that when the Gradient Boosting model is confronted with unseen test data, it achieves an accuracy which is slightly over 80 percent.

Keywords: *text classification, SSH, sociology, supervised machine learning, abstracts*

Conference Topic
Knowledge discovery and data mining

Introduction

Classifying scientific articles according to disciplines is most commonly done by making use of a proxy such as Clarivate Analytics' Web of Science (WoS) journal level Subject Categories (SC). Clarivate's staff working on WoS assigns the journals it indexes to one or more SCs and the publications that appear in these journals are treated as belonging to the same SC (for details on this procedure, see footnote 1 in Pudovkin and Garfield, 2002). Although this has proven to be useful, it has been treated as a limitation as well. Glänzel, Schubert, and Czerwon (1999), for example, point out that such an approach works well in the case of highly specialized journals, but that it is problematic for publications appearing in multidisciplinary or general journals. Solutions to this problem involve article level (re-)classifications based on the SCs of the references made in the article (cf. Glänzel et al., 1999), or clustering articles based on their citation relations (cf. Waltman and van Eck, 2012).

In the citation clustering study by Waltman and van Eck (2012) the authors distinguish between three different levels. At the highest level, the clusters correspond to 'broad scientific disciplines' (e.g. 'natural sciences', 'social sciences', etc.), and at the lowest level to 'small subfields'. This lowest level can be perceived of as research specialties, and consists of 22,412 clusters. Classifying publications on such levels of granularity is, as the authors acknowledge, difficult (Waltman and van Eck, 2012, p. 2386). Finding and manually assigning labels that adequately denote the clustered 'communities' on this scale becomes virtually impossible.

For publications in the social sciences and humanities (SSH) concerns regarding citation approaches can be further extended. First, the lack of coverage in indexing services like WoS poses additional problems. Far from all sources used in the SSH are covered by WoS or Scopus, which means that citation clustering will only yield a partial result. Second, when aiming for reference-based reclassification (see Glänzel et al., 1999), determining the SCs of the references of a publication is problematic, not only because of the lack of coverage of SSH journals, but also because SSH scholars tend to cite more books and non-source items (Ossenblok, Engels, and Sivertsen, 2012; Larivière, Archambault, Gingras, and Vignola-Gagné, 2006).

Making use of other sources or publication meta-data to classify publications, like for example author affiliations (cf. Guns, Sīle, Eykens, Verleysen, and Engels, 2018) might help to overcome these hurdles. But, as Guns et al. (2018) point out, author affiliations do not always correspond well with the cognitive domain SSH authors are working in. In times of increased specialization, moreover, departmental affiliations, like WoS SCs, are often too generic to get an adequate understanding of an author's expertise.

In this article we aim to counter the previously mentioned shortcomings by applying a text-based approach. We present the experimental results of a supervised machine learning (ML) exercise and assess its potential for a fine-grained automated classification of scientific articles in sociology.

The Flemish Discipline Code List (DCL) (2018)

The Flemish Discipline Code List (DCL) is used as our guiding classification scheme (Vancauwenbergh and Poelmans, 2018). The DCL is structured as a hierarchical tree with four levels. The first level refers to 7 broad fields of science. To allow for international comparison, the first level of the DCL conforms to the highest level of the OECD Fields of Science coding scheme (2007) (hereafter referred to as FOS). For the case of sociology and anthropology, on the top level of the FOS we find category 5 'social sciences' and a rather generic sub-category '5.4 Sociology and Anthropology'.

Hierarchical structure of OECD FOS (2007)
5. Social Sciences
    5.4 Sociology and Anthropology

Hierarchical structure of DCL (2018) (ex. "050405")
05 Social Sciences
    0504 Sociology and Anthropology
        050401 Anthropology (14 subcategories)
        050402 Applied Sociology (11 subcategories)
        …
        050405 Social Change (2 subcategories)
            05040501 Social change
            05040502 Social movements and collective action
            05040599 Social change not elsewhere classified
        …

Figure 1. Excerpt of tree structure: OECD FOS (2007) coding scheme and DCL (2018)

The third and fourth layer of the DCL add two more granular layers representing disciplinary subfields. The third layer might be interpreted as referring to subdisciplinary categories, whilst the fourth level can be considered as referring to research specialties, in this case within sociology. To construct and define this most granular level, experts from the corresponding fields were consulted. In total, the DCL contains 2,866 codes (for details, cf. Vancauwenbergh and Poelmans, 2018). Our

objective is to automatically classify articles (based on abstracts and titles) into the most granular categories of the coding scheme.

Methods

Since 30 years, the dominant paradigm of Text Classification (TC) consists of ML approaches. ML algorithms are deployed such that "*a general inductive process automatically builds an automatic text classifier by learning, from a set of preclassified documents, the characteristics of the categories [or labels] of interest*" (Sebastiani, 2002, p. 2). ML approaches have already been applied to classify abstracts of journal articles (also full-text or parts thereof) (for a recent example, see Langlois, Nie, Thomas, Hong, and Pluye, 2018). Our approach is different from previous studies as we select articles and abstracts from one social sciences discipline, sociology, and assess the accuracy of supervised ML algorithms for classifying these abstracts into 77 different sub-disciplines, or specialties.

Data collection

As noted, the development of supervised ML models involves collecting a set of preclassified documents; the training data. Based on these data, a model can be learned with which we can predict an output for a yet unclassified input. Put differently, the training data consists of examples of already classified input/output pairs (Müller and Guido, 2017). As we envisage a granular classification of abstracts in sociology, the *Sociological Abstracts* database (SA) was used to construct a training data set.

SA covers articles published in 1000+ journals in the field of sociology and related social sciences (i.e. anthropology, social psychology, etc.), dissertations, books, conference papers and proceedings dating back to 1952. The sources covered are relatively diverse, with over 40% of titles being published outside North America. SA was queried for abstracts fitting in the most granular categories of the DCL.

The main advantage of the service provided by SA, is the acknowledged Thesaurus of Sociological Indexing Terms (Booth, 1996; Blaemers, 2006). This thesaurus allows for structured and specific queries. Additionally, it is possible to determine publication dates, publication types, peer-review status and publication language. For this experiment, we have limited ourselves to peer-reviewed English language journal articles published in the period 2000-2018. To control for adequacy, the queries were manually performed between December 6th (2018) and December 24th (2018).

We sorted the results by relevance, visually inspected the top 100 results to further ensure the accuracy of the results, and downloaded the metadata for the first 1,000 articles. If the query returned less than 1,000 results, we downloaded the metadata of all articles. The metadata which we retrieved from SA include information on: publication title, author names, journal title, journal ISSN, full abstract, unique identifier assigned by SA, etc. The downloaded metadata were then labeled with the discipline classification codes. In accordance with the number of categories, 77 queries were conducted. This resulted in a dataset with 66,251 entries.

Data cleaning and processing

While exploring the data in preparation of processing, some anomalies were discovered. To account for these, abstracts with more than 600 or less than 60 words, as well as documents with a publication date before 2000, were omitted. After cleaning and labelling the data, we retained 48,961 labelled abstracts. On average, each DCL research specialty contains 635.86 labelled abstracts (min. = 143, max. = 924).

As ML algorithms typically demand for a numeric matrix as input, in our case, the 'features' or columns of this matrix are representations of the words in the abstracts of the publication. These features were first tokenized making use of natural language processing techniques. NLTK, a natural language toolkit implemented in Python, was used. First, we removed English language stop words and performed snowball stemming (cf. http://snowballstem.org/). Subsequently, the scores contained in the cells of the matrix were vectorised with TF-IDF. Different parameter settings for the ML models were tested making use of Hyperopt (Bergstra, Corner, Eliasmith, Yamins, and Cox, 2014).

Machine learning (ML)

ML approaches come in diverse forms. A distinction can be made between supervised and unsupervised approaches. In this experiment we make use of supervised learning, whereby the algorithm learns from a training data set of abstracts correctly labeled as belonging to a sub-discipline. In this experiment, we evaluate four different supervised ML algorithms, namely: Multinomial Naïve Bayes (MNB), Support Vector Machine (SVM), Random Forest Classifier (RFC), and Gradient Boosting (GB).

The four algorithms were chosen based on their popularity and proven success when implemented in similar scenarios. The implementation of the first three algorithms was carried out with the Scikit-learn package (version 0.20.1) in Python. Scikit-learn harnesses a broad set of ML algorithms (Pedregosa et al., 2011). Given its consistency and relative ease of use, it enables comparison of different applications. For gradient boosting we used the LightGBM implementation (Ke et al., 2017). Accuracy of the algorithms is measured when the models created by the algorithms are fitted to 'unseen' data (i.e. the test data). This score is calculated by dividing the number of correctly categorized documents by the total number of documents.

Preliminary results

Both MNB and SVM perform relatively poor when compared to the ensemble classifiers (i.e. RFC and LightGBM). The degree to which these latter two classifiers outperform the former is worth noticing. SVM outperforms MNB, but is considerably less accurate than RFC and LightGBM. Figure 2 presents a heat map of the results obtained by the LightGBM algorithm. The rows depict the true labels, and the columns depicts the labels predicted by the algorithm. The diagonal in this diagram represents the number of cases that were correctly assigned.

Algorithm	*Achieved accuracy on test set*
Multinomial Naïve Bayes (MNB)	0,49299
Support Vector Machine (SVM)	0,68346
Random Forest Classifier (RFC)	0,71306
Gradient Boosting (LightGBM)	0,81449

Table 1. Accuracy results for each classifier

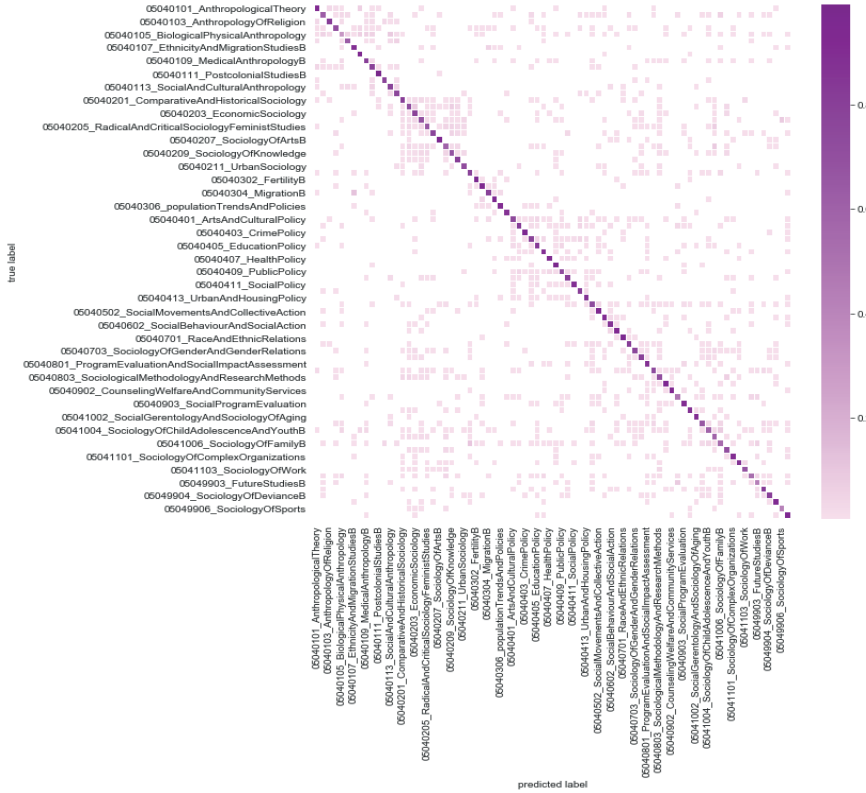


Figure 2. Heatmap of LightGBM results: predicted labels test set (20 % of data) in the columns, true labels in the rows

Conclusion

Our results show that LightGBM can be a fruitful approach to overcome difficulties with regard to a granular classification of scientific articles in SSH. The algorithm was able to correctly classify over 80% of the abstracts collected from SA. Our visualization (cf. figure 2) show that if mistakes were made, the publications were mostly assigned to neighboring specialties.

A distinction can be made between single- and multi-label TC. The former involves classifying each abstract into one discipline, while in the latter case abstracts may be classified as belonging to multiple disciplines. Whereas the latter would be a more natural way of approaching scientific abstracts (i.e. more often than not, in one single document, there exists a significant overlapping of topics or disciplinary perspectives thereon), in this experiment we focused on single label classification. An exploration of multi-label approaches would be appropriate.

Acknowledgements

This investigation has been made possible by the financial support of the Flemish Government to ECOOM, among others. The opinions in the paper are the authors' and not necessarily those of the government.

References

Bergstra, J., Komer, B., Eliasmith, C., Yamins, D., & Cox, D. D. (2015). Hyperopt: a python library for model selection and hyperparameter optimization. *Computational Science & Discovery, 8*(1),

Blaemers, J. (2006). A History of the Development and Maintenance of the Sociological Abstracts Thesaurus of Sociological Indexing Terms. *Behavioral & Social Sciences Librarian, 25*(1), 111-122. doi: https://doi.org/10.1300/J103v25n01_05

Booth, B. (1996). *Thesaurus of Sociological Indexing Terms*. Sociological Abstracts, Inc. The University of Michigan.

Glänzel, W., Schubert, A., & Czerwon, H. (1999). An item-by-item subject classification of papers published in multidisciplinary and general journals using reference analysis. *Scientometrics, 44*(3), 427-439. doi: https://doi.org/10.1007/BF02458488

Guns, R., Sīle, L., Eykens, J., Verleysen, F. & Engels, T. C. E. (2018). A comparison of cognitive and organizational classification of publications in the social sciences and humanities. *Scientometrics, 116*(2), 1093-1111. doi: https://doi.org/10.1007/s11192-018-2775-x

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q. & Liu, T. Y. (2017). Lightgbm: A highly efficient gradient boosting decision tree. In: *Advances in Neural Information Processing Systems*, 3146-3154.

Kibriya A.M., Frank E., Pfahringer B., Holmes G. (2004). *Multinomial Naive Bayes for Text Categorization Revisited*. In: Webb G.I., Yu X. (*eds*) AI 2004: Advances in Artificial Intelligence. AI 2004. Lecture Notes in Computer Science, 3339. Springer, Berlin, Heidelberg. doi: https://doi.org/10.1007/978-3-540-30549-1_43

Langlois, A., Nie, J., Thomas, J., Hong, Q. N., and Pluye, P. (2018). Discriminating between empirical studies and nonempirical works using automated text classification. *Research Synthesis Methods, 9*(4), 587-601. doi: https://doi.org/10.1002/jrsm.1317

Larivière, V., Archambault, É., Gingras, Y., & Vignola-Gagné, É. (2006). The place of serials in referencing practices: Comparing natural sciences and engineering with social sciences and humanities. *Journal of the American Society for Information Science and Technology, 57*(8), 997-1004. doi: https://doi.org/10.1002/asi.20349

Ossenblok, T. L., Engels, T. C., & Sivertsen, G. (2012). The representation of the social sciences and humanities in the Web of Science—a comparison of publication patterns and incentive structures in Flanders and Norway (2005–9). *Research Evaluation, 21*(4), 280-290. doi: https://doi.org/10.1093/reseval/rvs019

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V. Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M. Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research, 12*(2011), 2825-2830.

Pudovkin, A. & Garfield, E. (2002). Algorithmic procedure for finding semantically related journals. *Journal of the Association for Information Science and Technology, 53*(13), 1113-1119. doi: https://doi.org/10.1002/asi.10153

Sebastiani, F. (2002). Machine Learning in Automated Text Categorization. *ACM Computing Surveys, 34*(1), 1-47. doi: https://doi.org/10.1145/505282.505283

Suominen, A., & Toivanen, H. (2015). Map of science with topic modeling: Comparison of unsupervised learning and human-assigned subject classification. *Journal of the Association for Information Science and Technology, 67*(10), 2464-2476. doi: https://doi.org/10.1002/asi.23596

Vancauwenbergh, S., & Poelmans, H. (2018, June). *The creation of the Flemish research discipline list, a huge step forward in harmonizing research information (systems)*. Presented at CRIS2018: 14th international Conference on Current Research Information Systems, Umeå, Sweden.

Exploring the impact of scholarly journals in social sciences and humanities upon patentable technology

Felix de Moya-Anegon[1], Carmen Lopez-Illescas[2], Vicente Guerrero-Bote[3], Henk F. Moed[4]

[1]*felix.moya@scimago.es*
SCImago Group, Madrid, Spain

[2]*carmlopz@gmail.com*
Presenter, University Complutense of Madrid. Information Science Faculty. Dept. Information and Library Science, SCImago Group, Spain

[3]*guerrero@unex.es*
SCImago Group, Dept. Information and Communication, University of Extremadura, Badajoz, Spain

[4] *henk.moed@uniroma1.it*
Sapienza University of Rome, Italy

Abstract

This paper examines the impact of papers published in scientific-scholarly journals upon patentable technology as reflected in examiner- or inventor-given citations in patents to the scientific literature. It analyses data created by SCImago Research Group linking PATSTAT's scientific non-patent references (SNPRs) to source documents indexed in Scopus. The frequency of patent citations to journal papers is calculated per year, institutional sector, subject field and "top" journal. PATSTAT/Scopus-based statistics are compared to those derived from Web of Science/USPTO linkage. In addition, the current paper analyses the impact of research publications in selected journals covering four subject fields of social sciences and humanities (SS&H): informetrics/scientometrics, history of science, music and education. Looking behind the numbers, it presents typical examples of the titles of citing patents and cited papers, and word clouds of these titles. The analysed SS&H fields and journals do generate impact upon technology. It is proposed to further develop ways to combine impact numbers with citation content and context analysis.

Introduction

Brief literature review on the analysis of scientific non-patent references in patents

Anthony van Raan (2017) gives an excellent overview of main developments in the use of patent citations to the scientific literature, starting with the work of Francis Narin and co-workers who explored measures of the "science intensity" of technological fields and showed already in the 1990s how major inventions patented by industrial firms at the US Patent Office depend upon publicly funded basic research (e.g., Narin, Hamilton, & Olivastro, 1997). Van Raan used the acronym SNPRs to indicate Scientific Non-Patent References in patents. He concluded that "Only a small minority of publications covered by the Web of Science or Scopus are cited by patents, about 3-4 per cent. However, for publications based on university-industry collaboration the number of SNPRs is considerably higher, around 15%" (Van Raan, 2017, p 13). The studies on which this conclusion is based, were based on non-patent citations in USPTO (US Patent Office), linked with Web of Science or its predecessors (mainly the Science Citation Index).

Several studies showed that not only the distribution of SNPRs on the "cited side" among *target* WoS papers is skewed, but also - on the "citing side" - the distribution of SNPRs among *source* patents. Of course, these manifestations of skewness of citations on the citing and cited side have also been observed in the citation analysis of scientific papers (e.g., Price, 1965), but the skewness is much stronger for patent-to-paper citations than it is for paper-to-paper citations.

Other studies observed a national patent citation bias: patents submitted by applicants from a particular country showed a preference for citing research papers published by authors affiliated with institutions located in the same country. Other papers found a positive correlation between a country's technological performance and its scientific strength, and provided evidence that in *emerging fields* of technology the number of SNPRs in patents is higher than in other fields (Van Looy et al., 2006). In his review, Van Raan underlined that the number of SNPRs in patents, and the probability that a scientific publication may be used as an SNPR, depends on a series of factors including the stage of development of a technological field; the distribution of SNPRs among inventors and examiners; characteristics of the patent office, and of the applicant firms; and differences in the economic value of patents. He concluded that "SNPRs indeed form a bridge between science and technology, but more in a broader sense, i.e. at a macro-level such as the "science intensity" of technological fields or the science-technology interaction at the level of countries" (Van Raan, 2017, p. 22).
Defining the time lag between a scientific breakthrough and an invention as "the time lapse between the publication year of a paper and the year this paper is cited in a patent", Van Raan pointed out that large differences appear to exist in time lags between technological fields. He also warned that "the SNPRs may represent important recent scientific research but this research on its turn may be based on even more important, earlier breakthrough work, not cited in the patent but perhaps cited in the SNPRs."

Measuring the technological impact of scientific-scholarly subject fields

At the technology side, in the analysis of linkages between patent citations and scientific-scholarly papers, measures are calculated of the science intensity, science base or science linkage of (patentable) technology. Francis Narin and co-workers defined science linkage as "a measure of the extent to which a company's technology builds upon cutting edge scientific research. It is calculated on the basis of the average number of references on a company's patents to scientific papers, as distinct from references to previous patents. Companies whose patents cite a large number of scientific papers are assumed to be working closely with the latest scientific developments" (Narin, Breitzman & Thomas, 2004).
From the science side, patent citations can be used to calculate indicators of the technological impact of scientific-scholarly work. Such indicators aim to capture the extra-scientific or 'societal' impact of research. For instance, Halevi and Moed (2012) examined the impact of research published in library science journals upon technology as reflected in SNPRs, using TotalPatent™, a LexisNexis product linked with Scopus (TotalPatent, n.d.). A good overview of methodological approaches to statistical patent analysis is given in Hinze & Schmoch (2004).

Aim of the paper and research questions

The aim of this paper is first of all to give a comprehensive overview of the frequency at which patents processed in PATSTAT cite Scopus source articles. The research question addressed in the first part is: What is the percentage of papers cited in patents and the average number of patent citations per paper in the various main scientific-scholarly disciplines? And how does this frequency vary during the time period 2008-2017? A main objective is to compare the outcomes obtained in the PATSAT/Scopus database with those indicated in Van Raan's review, which are mostly based on linking USPTO NPRs to Web of Science. A breakdown is presented of the frequency of SNPRs by economic/institutional sector assigned in the SCImago database to the institutions of the authors of the cited papers, and by main subject category, using a classification implemented in Scopus of journals into 27 disciplines. The outcomes provide a statistical background for the analyses presented in the second part of this paper.

This second part presents a series of in-depth case studies of particular journals in selected subject fields (journal categories), especially but not exclusively in social sciences and humanities. It 'looks behind' pure citation numbers, by presenting lists of titles of citing patents and cited papers, and of the affiliations of the applicants. Both lists shed light upon the type of knowledge flow from science to technology that has taken place. It follows an analysis model applied in an earlier paper by Halevi and Moed (2012) on the technological impact of library science. This study found that research papers in library science had a considerable impact as reflected in patent citations. The current paper addresses in whether such an impact can also be found in other subject fields in social sciences and humanities (SS&H).

Data collection and handling

PATSTAT, "EPO worldwide PATENT STATistical Database", is a global patent database created by the EPO, published for the first time in 2008, to help patent statistical research at the request of a working group on patent statistics led by the Organization for Economic Cooperation and Development (OECD). Other members of this working group are: World Intellectual Property Organization (WIPO), Japan Patent Office (JPO), United States Patent and Trademark Office (USPTO), Korea Intellectual Property Office (KIPO), National Science Foundation of the United States (NSF) and European Commission (CE). As main advantages over other databases such as NBER (USA) or IIP (Japan) has worldwide coverage, the inclusion of more information and the existence of some auxiliary products that solve some of its problems, what has made it a de facto standard (Kang and Tarasconi, 2016). Its disadvantages are its orientation to Europe (data from national offices are exchanged with the EPO on the basis of agreements that change over time and may leave gaps) and its orientation to the review process (data that are not vital in the process of the patent examination has a lower quality).

PATSTAT is a relational database. It can be purchased on a DVD to be installed on a local computer or online, and can be consulted using SQL (De Rassenfosse, Dernis and Boedt, 2014). The EPO publishes two annual editions of PATSTAT, Spring and Autumn. The 2018 Spring Edition of PATSTAT (PATSTAT - Spring Edition of 2018) is a snapshot of the data present in DOCDB EPO, a global bibliographic database that includes data from more than 90 patent offices around the world, and the global database of legal information of INPADOC EPO. State, taken in the fifth week of 2018.

One of the PATSTAT tables includes the references to the non-patent literature. This table contains the full non-patent references, which do not follow a fixed format, and are not always complete. The table also contains a series of related data fields, but in many cases the values in these fields are missing. In a combined automated and manual approach, the records in this table were matched one by one against the source documents indexed in Scopus. This work was carried out by SCImago Research Group. These authors have designed a procedure divided in four phases (Guerrero-Bote, Sánchez-Jiménez & Moya-Anegón, 2019):

1. Data preprocessing: Preparation of data to facilitate and streamline subsequent processes. The most important actions: unify records; locate patterns corresponding to DOIs; assign publication years; normalize lexical variants and eliminate special characters; locate possible elements of the reference: first author, title, source; generate an inverted index with extracted terms.

2. Pre-selection of candidate couples. With the previous data of the preprocess, we have 9x1014 possible pairs formed by a NPL reference of PATSTAT and a reference of Scopus. Due to the lack of standardization, a direct comparison is necessary and that is impossible to address in such a large number of couples. For that reason, this phase aims to reduce that number to a sufficiently large number to minimize the possibility of a real couple being left

out. • To this end, a series of rules are used that are applied in the form of SQL statements in the data obtained from the previous phase.
3. Automatic evaluation of the candidate couples. The objective of this phase is to assign a score that allows to select for each NPL reference the Scopus reference that probably refers to the same document. For this purpose, a series of routines have been designed that look for the most important elements of the Scopus reference in the record of the TLS214_NPL_PUBLN table. The overall score is obtained by the product of the scores obtained for each element of the reference (by way of probability).
4. Human validation. An NPL reference may not have a Scopus candidate reference, it may have one or it may have several. Logically, if any of the candidates corresponds to the NPL reference, this should be the highest score obtained, but it is possible none of the assigned ones was valid. For this reason, a manual validation is necessary. To this end, an application has been developed that allows the cooperation of many people in human validation.
A patent family can be defined as "a set of patents taken in various countries to protect a single invention (when a first application in a country – the priority – is then extended to other offices)." In other words, a patent family is "the same invention disclosed by a common inventor(s) and patented in more than one country ("Patent families", n.d.)". One of the problems in patent citation analysis is that there may be substantial differences between members of the same family as regards the non-patent references they may contain. This may be especially the case for the examiner-given references. Patents of a family tend to pass a different evaluation process in each office. In some cases this process is faster and in others slower, and some they incorporate more non-patent literature references than in others. To avoid these differences, SCImago Research Group has retrospectively assigned all non-patent references in the various members of a family to each patent in that family. In this way, when a patent is granted, it incorporates all scientific non-patent references in its entire family.

Results

Analysis by institutional sector

Table 1 presents a breakdown of papers and patent citations by institutional sector of the papers. Data relate to the time period 2008-2017. A paper may be assigned to multiple sectors, if it results from a collaboration between authors active in different sectors. Therefore, the total number of assignments exceeds the number of papers by some 15 percent. For all sectors combined, the share of papers cited in patents is 3.2 % if double counts due to these multiple assignments are included, and 2.7 % otherwise. The largest percentage of papers from a particular sector cited in patents, relative to the total number of papers assigned to that sector, is the private sector (7.9 %), followed by the health sector (4.2 %). The share of private sector papers relative to the total number of papers in all sectors is only 2.6 %, but its share relative to the total number of received patent citations is 11 %, similar to that for the government sector.

Table 1. Number of papers and patent citations per papers' institutional sector (time period 2008-2017)

Papers Institutional Sector	*Papers*		*Citations in Patents to Papers*		*# Papers Cited in Patents*		*% Papers Cited in Patents*
	N	*%*	*N*	*%*	*N*	*%*	
Higher Education	18,534,000	68.1%	4,129,000	53.0%	527,000	60.1%	2.8
Government	4,137,000	15.2%	930,000	11.9%	134,000	15.3%	3.2
Health	3,565,000	13.1%	1,811,000	23.2%	150,000	17.1%	4.2
Private	719,000	2.6%	854,000	11.0%	57,000	6.5%	7.9

Other	250,000	0.9%	66,000	0.8%	8,000	0.9%	3.2
Total paper-sector assignments	27,205,000	100.0%	7,790,000	100.0%	876,000	100.0%	3.2
Total unique papers (excl. 'double counts')	23,511,000		5,351,000		628,000		2.7

Van Raan (2017) indicated in his review an overall percentage of papers cited in patents of 3-4, obtained in studies that were mostly based on USPTO and WoS. In the current study, based on Scopus and PATSTAT, this percentage is somewhat lower. Table 1 also shows that the percentage of cited papers in patents (last column) is for *paper-sector assignments* (semi-last row, 3.2 %) somewhat larger than that for total number of *unique* papers (bottom row, 2.7%). This means that multi-sector papers tend to attract somewhat more patent citations than single-sector publications do. Data for collaborative papers between the public and the private sector are not available in the current study. Therefore, van Raan's conclusion that papers co-published between the public and private sector have a relatively large percentage of papers cited in patents, cannot be directly validated.

Overall results per year

Table 2 presents the annual trend in the number and percentage of papers cited in patents during the 10-year period 2008-2017. A 5-year citation window is used. This means that only citations are counted to publications published during the *five* years preceding the filing year of a patent.

Table 2. Annual trend in the number and percentage of papers cited in patents between 2008-2017

Year	Nr Journals (sourceids) with at least 1 citable doc in Scopus in past 5 years	Nr journals receiving at least 1 patent citation to citable docs in past 5 yrs	% journals receiving at least 1 patent citation to citable docs in past 5 yrs	Number of citable docs in 5 preceding years in Scopus	Number of citable docs from past 5 years cited in at least one patent	% citable docs from past 5 years cited in at least one patent	Total patent citations to citable docs from past 5 years	Number of patent citations per citable doc (5yrs)	Total citing patent families to citable docs from past 5 years	Number of citing patent families per citable doc (5yrs)
2008	23,820	8,424	35.4%	7,847,445	176,321	2.25%	598,989	0.076	248,725	0.032
2009	25,368	8,779	34.6%	8,325,768	182,296	2.19%	600,541	0.072	259,075	0.031
2010	27,240	9,349	34.3%	8,825,158	188,379	2.13%	607,578	0.069	266,009	0.030
2011	28,716	9,856	34.3%	9,291,259	191,230	2.06%	609,970	0.066	272,490	0.029
2012	30,094	10,305	34.2%	9,783,094	197,467	2.02%	639,670	0.065	278,183	0.028
2013	30,758	10,727	34.9%	10,250,199	196,259	1.91%	600,011	0.059	272,744	0.027
2014	31,203	10,607	34.0%	10,722,918	185,699	1.73%	575,083	0.054	252,377	0.024
2015	31,890	9,940	31.2%	11,203,655	155,557	1.39%	496,650	0.044	207,536	0.019
2016	32,313	9,015	27.9%	11,575,432	119,278	1.03%	382,423	0.033	157,430	0.014
2017	32,669	7,697	23.6%	11,879,030	79,459	0.67%	245,469	0.021	101,388	0.009

Table 1 above revealed an overall percentage of papers cited in patents of around 3 per cent. The percentages in Table 2 are lower, namely 1-2 %. The difference is explained by the fact that the percentages presented in Tables 1 and 2 are based on different citation windows: Table 1 includes all citations and papers from a ten-year period, while Table 2 applies a 5-year citation window. The decline in the number of applications with at least one patent citation as from 2008, and especially during 2015-2017, is difficult to interpret. The 5-year publication period of cited papers is defined relative to the *filing* (or application) year, *not* relative to the *granted* year. As a consequence, examiner- given cited references published in or after the filing or application year are *not* included in the patent citation counts. If the time

period of patent examination is long, say, for instance, 5 years, the number of these non-counted cited references may be substantial. It is therefore hypothesized that the declining trend in the number of patents with one or more patent citations observed in Table 2 is at least partly due to the fact that the duration of the global patent granting process strongly increases in recent years, which on its turn may be caused by a strong increase in the number of submitted patent applications. The latter increase is not (yet) visible in the data, because the counts relate only to patents that were actually granted before 2018.

Analyses by subject category

Figure 1 compares for each of about 300 Scopus journal categories a standard *journal* impact factor-like measure (denoted with the acronym JIF) with one based on *patent* citations. The horizontal axis displays the mean number of *paper* citations per article calculated for each journal covering a particular Scopus journal category, applying a three-year window, counting, for instance, citations in 2012 to citable documents published in the three preceding years 2009-2011.

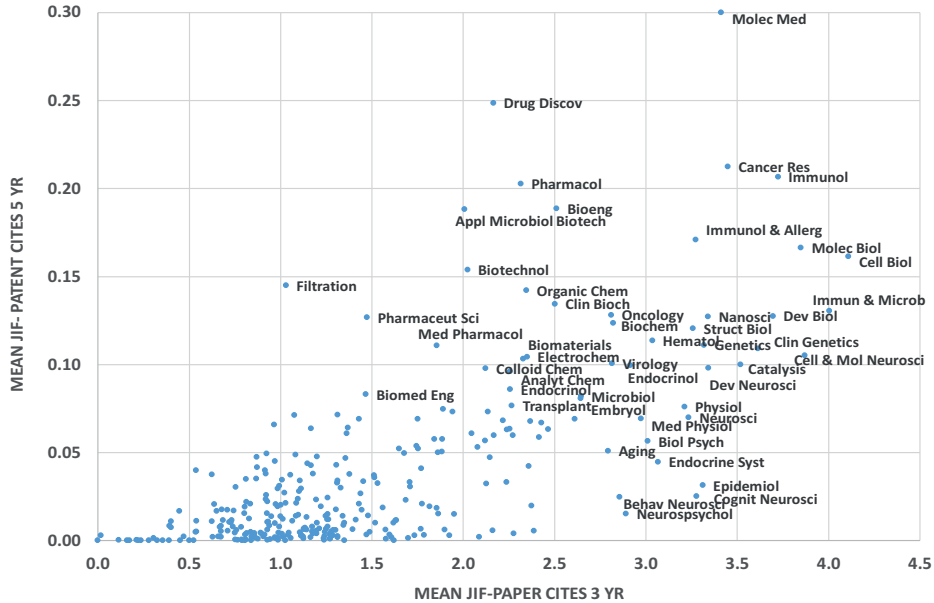


Figure 1. Mean journal impact factors based on paper citations and patent citations, per journal category, and for the (first filing) year 2012.

The vertical axis gives a similar measure, but now based on *patent* citations, and applying a 5-year rather than a 3-year window. Medical-biological categories show on average the largest patent-based JIFs. At this level, relative differences between categories hardly change if one analyses *patent family* citations instead of *patent* citations. The mean citation rates for these two types of citations per category show a very strong linear correlation: R-square is 0.97.

Analyses by journal

The paper- and patent-based "journal impact factor" (JIF) displayed in Figure 1 are also calculated in columns 5 and 7 in Table 3 below, but here at the level of *individual* journals instead of journal *categories*. Table 3 lists the 10 journals with the largest impact score based on patent citations for a single year: 2012. As indicated in Table 2 above, at the level of *all*

journals in all fields combined, and for the year 2012, the average number of patent citations per citable document is 0.065. The level of the patent citation impact of the *top journals* in Table 3 is in the same order of magnitude as the paper citation-based impact factor of a median journal in many Scopus categories. Table 3 also shows in column 6 a citation impact measure based on (citing) patent families rather than citing patents. Roughly speaking, its scores are one-third to one-half of those based on all patent citations.

Table 3. The 10 journals with the largest percentage of 2007-2011 papers cited in 2012 patents

Rank	Journal Title	Nr citable docs in 2007-2011	% citable docs 2007-2011 cited in 2012- patents to	JIF based on number of patent citations and 5-year window	JIF based on number of patent *family* citations and 5-year window	JIF based on paper citations and 3-year window	SJR (Scimago Journal Rank) 2012 based on paper citations
1	Annual Review of Immunology	119	42.9	2.21	0.95	57.5	38.09
2	mAbs (covering Antibody R&D)	196	37.2	2.45	0.78	4.8	1.40
3	Advanced Drug Delivery Reviews	576	35.6	2.33	0.97	15.4	4.50
4	Progress in Polymer Science	214	35.5	1.60	0.95	31.1	10.00
5	Pharmacological Reviews	107	31.8	1.81	0.57	24.8	10.67
6	Journal of Medicinal Chemistry	3,907	31.3	1.68	0.52	5.9	2.34
7	Proc Int. Symp. on Computer Architecture	210	31.0	0.93	0.41	10.2	2.33
8	Nature Biotechnology	988	30.9	3.85	1.33	19.1	10.87
9	Metabolic Engineering	260	30.4	1.44	0.61	7.5	2.99
10	Molecular Therapy	1,194	30.3	1.12	0.36	7.9	3.23

Results for selected journals from social sciences and humanities

While Halevi and Moed (2012) proved in their work the economical and technological impact of library science, the current study aims to explore whether such influence can also be found in other Social Sciences and in Humanities' subject fields. To that end, three typical journal categories from the field were chosen: History and Philosophy of Science, Music, and Education. Also Library and Information Science was included in order to compare results with those obtained by Halevi and Moed (2012). Data in the current work relate to patents processed in PATSTAT granted during the time period 2008-2017 and citations to Scopus articles published in the five years preceding the patent's filing year. Table 4 presents for 12 journals in four SSH subject fields typical examples of titles of citing patents and cited papers. In addition and also with the purpose of gaining insight into the specific subjects of the citing patents and cited papers, word clouds are presented using the Worditout software (http://worditout.com) for two subject fields: Library & Information Science (Figures 2a and 2b) and Education (figures 3a and 3b).

In library science the research cited in patents, according to the results presented by Halevi and Moed (2012), were mostly featuring library information and customer management systems together with classification and indexing methodologies. The patents citing these articles were found to feature mainly online commerce applications. In the current work we can also observe the important role played by these subjects in the titles of citing patents as well as in cited papers. In the world clouds for Library & Information Science (Fig. 2a and 2b) it can be also detected the importance of this web based technology in the field by the size of words such as *web, google, machine, technologies, computing, engine, link, databases,*

automated, multimedia, browsing, network, site or XML, which reflect their frequency. However and not surprisingly, the larger and bolder words displayed in the word clouds for papers and patents in Scientometrics are related to information and evaluation. Among these words we find: *measures, search, analysis, similarity, data, document, information, content, knowledge, journal, performance, survey, table, evaluating, factorizing, matrix, graph, weighted, comparison, results, bibliometric, citation, clusters, quality, mapping or ranking.*

Table 4. Examples of patent citations to selected SS&H journals

Field / Journal	*Citing patent title (Example)*	*Cited paper title (Example)*
Library & Information Science		
Journal of Information Science (56/15)	System and method for annotating documents	Usage patterns of collaborative tagging systems
	Method and device for enriching a content defined by a timeline and by a chronological text description	Aspect-based sentiment analysis of movie reviews on discussion boards
	Systems and methods of handling internet spiders	Web robot detection in the scholarly information environment
Journal of Informetrics (6/3)	Incorporating lexicon knowledge into svm learning to improve sentiment classification	Sentiment analysis: A combined approach
	Systems and methods for ranking nodes of a graph using random parameters	Finding scientific gems with Google's PageRank algorithm
	System and method for analyzing and categorizing text	Applying social bookmarking data to evaluate journal usage
Scientometrics (18/14)	Apparatus and method for determining stage using technology lifecycle	Anticipating technological breakthroughs: Using bibliographic coupling to explore the nanotubes paradigm
	System method and program to test a web site	'Mini small worlds' of shortest link paths crossing domain boundaries in an academic Web space
	Social network model for semantic processing	A quantitative analysis of indicators of scientific performance
History & Philosophy of Science		
British Journal for the Philosophy of Science (15/2)	Diagnosability system: flood control	Causes and explanations: A structural-model approach. Part I: Causes
IEEE Annals of the History of Computing (17/5)	Desktop stream-based information management system	Alan Kay: Transforming the computer into a communication medium
	Method and system for processing email attachments	The technical development of internet email
	Ambient backscatter transceivers apparatuses systems and methods for communicating using backscatter of ambient rf signals	Implications of historical trends in the electrical efficiency of computing
Philosophy, Ethics, and Humanities in Medicine (54/2)	Composition and methods of use for treatment of mammalian diseases	Are animal models predictive for humans?
	Novel biomarkers	Rethinking psychiatry with OMICS science in the age of personalized P5 medicine: Ready for psychiatome?
Music		
Computer Music Journal (58/12)	Apparatus and method for enhanced spatial audio object coding	The spatial sound description interchange format: Principles, specification, and examples
	Method and system for extracting tempo information of audio signal from an encoded bit-stream and estimating perceptually salient tempo of audio signal	Complexity-scalable beat detection with MP3 audio bitstreams

	System and method for performing automatic audio production using semantic data	An offline, automatic mixing method for live music, incorporating multiple sources, loudspeakers, and room effects
Music Perception (26/8)	Personalized auditory-somatosensory stimulation to treat tinnitus	Listening to filtered music as a treatment option for tinnitus: A review
	Systems methods and media for identifying matching audio	Perceptual smoothness of tempo in expressively performed music
	Methods and devices for treating hypertension	Music and autonomic nervous system (dys)function
Organised Sound (13/2)	Apparatus and method for efficient object metadata coding multi-channel audio signals	Object-based audio reproduction and the audio scene description format
	Wearable sound	Imposing a networked vibrotactile communication system for improvisational suggestion
Psychology of Music (18/3)	Side effect ameliorating combination therapeutic products and systems	Exposure to music and cognitive performance: tests of children and adults
	A media player	Toward a better understanding of the relation between music preference, listening behavior, and personality
	Systems and techniques for identifying and exploiting relationships between media consumption and health	Listening to sad music in adverse situations: How music selection strategies relate to self-regulatory goals, listening effects, and mood enhancement
Education		
Computers and Education (74/29)	Disease therapy game technology	Exploring the potential of computer and video games for health and physical education: A literature review
	Personalizing error messages based on user learning styles	E-Learning personalization based on hybrid recommendation strategy and learning style identification
	Individual learning device and method based on radio communication network	The design and evaluation of a computerized adaptive test on mobile devices
Proceedings - Frontiers in Education Conference (28/11)	Hierarchical state machine generation for interaction management using goal specifications	Invited panel - Engineering technology education in the era of globalization
	Cloud desktop system with multi-touch capabilities	Web-based tools to sustain the motivation of students in distance education
	Tagging method	Web editing module for tagging metadata of the Fedora commons repository learning objects under DRD and LOM standards

Numbers between parentheses give the number of patent citations and the number of cited papers per journal

Method and *system* have not been included in the previous enumeration despite being the most outstanding words in the word cloud for patents since these are the evident words, for patenting nature reasons, occupying the first frequency positions in the other disciplines. The current results also coincide with Halevi and Moed's as regards the prominent role of indexing and classification methodologies in library science. What it is important to notice in the current picture is the numerous and highlighted words related to indexing and classification connected to the web based technologies. In the word clouds we find: *classification, social, tag, tagging, bookmarking, folksonomy, collaborative, community, index, ontology or semantic*. The technological influence of the research carried out in indexing and classification in relation to the social and semantic web is even more evident in

the papers and patents' titles: *collaborative tagging, sentiment analysis, web robot detection, Google's PageRank, social bookmarking, social network, semantic processing*, among others.

Figure 2a. Word Clouds for Library & Information Science based on citing patent titles

Figure 2b. Word Clouds for Library & Information Science based on cited paper titles

Figure 3a. Word Clouds for Education based on citing patent titles

Figure 3b. Word Clouds for Education based on cited paper titles

From the Web 1.0 to the challenging Web 4.0, from document sharing to data sharing library science has always been on the basis of the information science-technology interaction. The relationship between the development of library science and innovation in Web technology is not new but these results confirm the increasing importance of the technological impact of the subject field, since the expanding role of these technologies in redefining today's society. The upcoming trends in the fields of information technologies reveal the world as a globalized highly intelligent information space, where the individual needs will be socially customized.
This combination of social and personalization aspects is also present in the results obtained from the other studied SS&H categories. Words related to both customization and social issues are present in the word clouds as well as in the patent and papers' titles. We find the words *collaborative, social, personalizing, personal, community, personalized or individual* appear with a high frequency. In the other three analysed fields the research cited in patents were mostly featuring Web and electronic systems. For Education we find the words: *e-learning, online, electronic, virtual, web, web-based, computer-based, software, network, digital, computer-readable, urls, automatically, wireless, programming, remote, server*. In Music, *automatic* is one of the most highlighted words, also *computed, computer, digital and offline*. In History the most visible word is *computing*. Computing is one of the most common connections to other fields found in the SS&H research cited in patents, but not the only one. We find connection to Health in all categories except for Library & Information Science. In Education, Music and even more in History & Philosophy of Science we find words standing out such as *health, therapeutic, disease, medicine, diagnostic, cancer, antigen, cells or virus.*

Table 5. Most important assignees per research field

Library & Information Science		*History & Philosophy of Science*		*Music*		*Education*	
Assignee	*# Patents*	*Assignee*	*# Patents*	*Assignee*	*# Patents*	*Assignee*	*# Patents*
IBM	13	ORACLE	15	FRAUNHOFER	9	IBM	5
PALO ALTO RES. CENTER (Xerox)	7	BOEHRINGER INGELHEIM	6	THE INVENTION SCIENCE FUND I	8	MICROSOFT CORPORATION	4
MICROSOFT CORPORATION	6	UNIVERSITY OF PENNSYLVANIA	5	DOLBY INTERNATIONAL	5	MICROSOFT TECH LICENSING	4
THOMSON LICENSING	5	IBM	3	X-SYSTEM	5	SAMSUNG ELECTRONICS	4
						MINAPSYS SOFTWARE CORP	3
						MOTOROLA	3

Table 5 gives a list of the most important patent assignees in the four subject fields, including large technology firms, but also an investment company (Invention Science Fund I), a university (Univ Pennsylvania) and a research organization (Fraunhofer).

Conclusions

The study has shown that the analysed SS&H fields and journals do generate impact upon patentable technology. As far as could be analysed, the frequency of patent citation to the scientific literature based on PATSTAT/Scopus is in the same order of magnitude as that based on Web of Science/USPTO, but slightly lower. It is as of yet uncertain whether this difference is due to differences in patent source (PATSTAT vs USPTO) or to differences in publication source (Scopus vs Web of Science). A more systematic study of a wide range of journals and fields could generate relevant information in an assessment of the technological impact of scientific-scholarly research in general, and of particular fields under assessment. More advanced indictors of technological impact of scientific journals could be calculated. In order to be practically useful and convincing to users of patent citation-based impact indicator, it is proposed to deliver not merely numbers, but also provide information on the citation content, in terms of what is being cited, and in which context. World profiles of patent and paper titles presented in the current paper are useful tools but constitute a first step. More advanced mapping of citing and cited works is feasible and should be further explored.

References

De Rassenfosse, G., Dernis, H. and Boedt, G. (2014), An Introduction to the Patstat Database with Example Queries. Australian Economic Review, 47: 395-408. doi:10.1111/1467-8462.12073

European Patent Office (2018), Data Catalog PATSTAT Global. Versión 5.11. EPO PATSTAT.

Guerrero-Bote, V.P., Sánchez-Jiménez, R. & Moya-Anegón, F. (2019), The citation from patents to scientific output revisited: A new approach to the matching Patstat/Scopus. El profesional de la información (in press).

Halevi, G., & Moed, H.F. (2012). The technological impact of library science research: A patent analysis. In E. Archambault, Y. Gingras, & V. Larivière (Eds.), Proceedings of 17th International Conference on Science and Technology Indicators (Vol.1, pp. 371–380), Montréal: Science-Metrix and OST.

Hinze, S., & Schmoch, U. (2004). Opening the black box. In: Moed, H.F., Glänzel, W., and Schmoch, U. (2004) (eds.). Handbook of quantitative science and technology research. The use of publication and patent statistics in studies of S&T systems. Dordrecht (the Netherlands): Kluwer Academic Publishers, 215-235.

Kang, B., Tarasconi, G. (2016), PATSTAT revisited: Suggestions for better usage. World Patent Information, 46, 56-63.

"Patent families" (n.d.) https://en.wikipedia.org/wiki/Patent_family.

"TotalPatent" (n.d.). Lexis Nexis IP. https://www.lexisnexisip.com/products/total-patent-one/ .

Narin, F., Hamilton, K.S., and Olivastro, D. (1997). The increasing linkage between US technology and public science. Research Policy, 26, 317–330.

Narin, F., Breitzman, A. & Thomas, P. (2004). Using Patent Citation Indicators to Manage a Stock Portfolio. In W. Glänzel, H.F. Moed & U. Schmoch (Eds.), Handbook of Quantitative Science and Technology Research. The Use of Publication and Patent Statistics in Studies of S&T Systems (pp. 553-568). Dordrecht: Kluwer Academic Publishers.

Price, D.J.D (1965). Networks of scientific papers. Science, 149, 510-515.

Van Looy, B., Debackere, K., Callaert, J., Tijssen, R., & van Leeuwen, T. (2006). Scientific capabilities and technological performance: An exploration of emerging industrial relevant research domains. *Scientometrics*, 66(2), 295–310.

Van Raan, A.F.J (2017). Patent Citations Analysis and Its Value in Research Evaluation: A Review and a New Approach to Map Technology-relevant Research. *Journal of Data and Information Science,* 2, 13-50.

Democracy, Globalization, and Science

Travis A. Whetsell[1], Koen Jonkers[2], Ana-Maria Dimand[1], Jeroen Baas[3], Caroline S. Wagner[4]

[1] *Travis.whetsell@FIU.edu; Adima007@FIU.edu*
Florida International University, School of International and Public Affairs, Department of Public Policy & Administration, Miami, FL, USA, 33155
[2] *Koen.JONKERS@ec.europa.eu*
Knowledge for Growth, Finance and Innovation Unit, Joint Research Centre (JRC)[1], Brussels, Belgium
[3] *J.Baas@elsevier.com*
Elsevier B.V., Amsterdam, Netherlands
[4] *Wagner.911@OSU.edu*
The Ohio State University, John Glenn College of Public Affairs, Columbus, OH, USA 43210

Abstract
The relationship between democracy and science has long been discussed by scholars. However, the question has not been adequately addressed empirically at the international level. Using panel data on 161 countries over a seven-year period, we estimate the effects of democracy on scientific influence, controlling for globalization, population size, publication output, and level of international research collaboration. The results show that countries with higher levels of democracy tend to have higher impact science.

Introduction

Studies of science policy have long suggested that political, economic, and scientific systems are interdependent, but the nature of the relationship and the direction of causality remain unclear. Jasanoff (2011) noted that political culture is linked to scientific culture, with the two co-producing a system of knowledge creation. Stirling et al. (2018) noted that social progress owes a great deal to public support for science and technology; there is a broadening political acceptance of this view worldwide. However, empirical studies have yet to adequately address questions about the relationship between forms of government and the performance of national systems of science.

At the macro-level, the relationship between democracy and science has mostly been examined conceptually and with the use of exemplar cases, e.g. Merton (1973), Popper (1966), and Kitcher (2003). For instance, Popper (1966) argued that the open society permits greater contact with new ideas from abroad and those that arise internally. Recent research on international scientific openness and mobility also suggest a positive effect on national systems of science (Wagner et al. 2018a, 2018b). However, the relation between varieties of democracy, openness, and science remains unclear (Gao et al. 2017). This paper follows on research about openness, mobility and scientific impact, by comparing national measures of scientific influence to measures of governance and globalization.

This research note addresses the general research question, what is the relationship between democracy, globalization, and the impact of scientific papers among a large sample of 161 countries in the international system. We combine longitudinal data from the Economist Intelligence Unit's Democracy Index from 2008 to 2015 with data from Scopus on the production, collaboration, and impact of scientific research, and from the KOF globalization index, aggregated at the country level. We find strong correlations for democracy, globalization, and scientific impact. Results show that political culture has the strongest relative estimate

This article does not necessarily represent the official views of the European Commission. The European Commission nor anyone acting on its behalf can be held responsible for any use of the data or analysis contained herein.

among the democracy sub-components. We also find consistent positive results for globalization on science.

Literature

Wiesner et al. (2018) suggest that democracies may permit greater structural complexity and stability. Since science operates as a complex, loosely-coupled system with an emergent networked architecture (Wagner, 2008; Simon 1996), the mutual complexity of democracy and science may produce benefits for both. In addition to providing increased mobility and openness, the decentralization of political power in democracies may provide structural conditions for the emergence of complex interactions needed in science, technology, and innovation.

Merton (1973) suggests that democratic forms of government and the enterprise of science are compatible because of the shared feature of universalism. Merton defines the concept of universalism as "*preestablished impersonal criteria*: consonant with observation and with previously confirmed knowledge" (Merton, 1973, p. 270, emphasis in original). Merton (1973) suggests: "The imperative of universalism is rooted deep in the impersonal character of science", which clashes with discrimination based on personal or group identities. In a sense, democracies seek to decentralize power to the individual rather than centralizing it in a ruling class. However, Merton is unclear whether science itself should be democratic or whether science operates better within a democracy. In contrast to democratic governments, Merton (1973) suggests autocratic governments impose centralized control and social divisions that infringe on free inquiry: "In modern totalitarian society, anti-rationalism and the centralization of institutional control both serve to limit the scope provided for scientific activity" (Merton 1973, p. 278). However, Mokyr (2017) suggests democratic societies often constrain science to obtain certain outcomes.

As Popper (1966) suggested, closed autocratic societies tend to be philosophically situated on fixed but fragile historicist visions, which science threatens to undermine. Antiquated scientific conceptions emphasized "methodological essentialism" that attempted to identify unchanging Platonic forms underlying empirical phenomena. While compatible with an autocratic Republic envisioned by Plato, contemporary science challenges these fixed historicist visions. As the argument goes, scientists must be free to pursue inquiry, and open societies provide a better context for the flourishing of scientific activity.

The argument for the democracy-science connection may be extended through an analysis of structural complexity (e.g. Wiesner et al. 2018), where the organizational/institutional structure of science is critical to progress. The contemporary structure of science is decentralized in a modular global network with self-similar structure up and down levels of analysis (Wagner & Leydesdorff, 2005; Wagner, Whetsell, & Leydesdorff, 2017). As Simon (1996) observed, modular loosely coupled hierarchies can be more efficient at handling uncertainty than fully centralized hierarchies. Kontopoulos (2006) refers to this type of structure as a heterarchy.

The democracy-science compatibility thesis, however, is not given among scholars of science and technology studies. Nahuis & Van Lente (2008) suggests the direction of causality might be reversed. In this view, it isn't that democracy enhances science, but that science may undermine or displace democracy. Studies of science and technology have examined public participation in science (Lengwiler 2008), responsibilities of science given a democratic context (Stilgoe, Owen, & Macnagthen, 2013), and the potential dangers of science in undermining democracy (Durant, 2011). This is a sensible concern given the potential dangers that science poses to human subjects, with numerous examples of "administrative evil" unfolding in the name of scientific progress (Adams & Balfour, 1998). However, in a similar sense, democracy may also pose a unique threat to scientific innovation.

Gao et al. (2017) recently tested the hypothesis of whether democracy is conducive to technological innovation. They were unable to show such a relationship between democracy measures and patent data in their analysis. They do however confirm an emerging body of literature that suggests openness can be a strong factor in scientific and technological progress. Indeed, national openness appears to be strongly associated with scientific impact, where *openness* is operationalized through international co-authorship data and mobility statistics on the scientific workforce (Wagner et al. 2018a, 2018b; Chinchilla-Rodríguez 2018; Robinson-Garcia et al. 2019).

The degree to which a society is subjected to new ideas does not only depend on the extent to which indigenous flowers are allowed to bloom. New ideas, technologies and values may also originate from beyond the national context. Contacts with other cultures through trade and cultural exchange have been at the basis for many scientific and technological advances over the course of the past centuries and the degree of receptiveness to foreign ideas may explain a large share of the difference in long term scientific and technological performance between countries over time (Mokyr, 2017). Taylor (2016) argued that the extent to which societies are integrated in the global system measured in terms of trade, investments, culture and political ties following the KOF Globalisation index (Dreher, 2006; Gygli et al., 2018) could explain a relatively large degree of the variation in technological performance of countries. This article will also conceptualize openness in terms of the KOF measures of integration in the world economy.

Research Question: What is the relationship between democratic governance, globalization, and scientific research at global scale?

Methods

Data: We use the Economist Intelligence Unit Democracy Index data (2008-2015), which quantifies the level of democracy for 167 nations, 2009 is missing. As a robustness check, the Democracy Index results are compared with Polity IV data (Marshall, Jaggers, & Gurr, 2002). These approaches are both recognized as important measures of democracy (Green & Gallery 2015). Elsevier ‘s bibliometric data are used for measures of science, aggregated at the national level. The KOF globalization index was used for globalization (Dreher, 2006; Gygli et al., 2018). World Bank Open Data were used for population size.

Variables: The democracy index (DemIndex) is a composite index of the five sub-components, which are derived from sixty measures. The components include, functioning of government, electoral process and pluralism, political participation, civil liberties, and political culture. All are on a scale from 0-10. As a robustness check, two measures from Polity IV were used (Polity2 and Democ): Democ is a measure of democracy, 0-10 scale; and Polity2 includes democracy and autocracy, -10 to 10 scale. The KOF globalization index (GlobIndex) is a composite of numerous measures with three sub-factors: economic, social, and political. Control variables include countries’ percentage of papers that are international (PercInternational), a fractional count of the number of papers produced by the country, which has been standardized (NumPubs); and the log of the population size (Ln(Population)). Per capita GDP was considered, following Gao et al. (2018), but the measure is highly correlated with GlobIndex. For scientific impact, we use Elsevier’s fractional field weighted citation index (FracFWCI), used previously in Wagner et al. (2018a, 2018b). For this variable proportions of authors are summed when calculating volumes and are weighed when relative citation impact is aggregated to the total citation impact of an entity (Waltman & van Eck 2015).

Methods: The analysis begins with Pearson correlation. Then cross-sectional regression is used with standardized coefficients and heteroscedastic robust standard errors and t-statistics (2008-2015). Then panel regression techniques are used (2010- 2015), including two-way

random effects and two-way fixed effects with robust panel clustered standard errors to account for heteroscedasticity and serial correlation over time.

Results

Table 1 presents descriptive statistics and the Pearson correlation matrix of variables used in the analysis. Uneven data coverage in the data sets across the period results in different sample sizes for the variables. The analysis shows that the DemIndex and its sub-components, as well as Glob.Index, are correlated with FracFWCI, warranting further analysis.

Table 1 – Descriptive Statistics and Pearson Correlation Matrix

	N	Mean	Std Dev	1	2	3	4	5	6	7	8	9	10
1-Dem.Index	1169	5.52	2.20	1									
2-Pol.Culture	995	5.57	1.66	0.73	1								
3-Pol.Participation	995	4.87	1.88	0.83	0.61	1							
4-Funct.Government	995	4.96	2.48	0.91	0.70	0.69	1						
5-Elect.Proc.Plural	995	5.97	3.54	0.93	0.51	0.72	0.78	1					
6-Civ.Liberties	995	6.21	2.76	0.95	0.59	0.72	0.83	0.91	1				
7-Glob.Index	1127	62.03	14.46	0.76	0.59	0.65	0.73	0.64	0.70	1			
8-Perc.International	1169	0.40	0.22	-0.34	-0.23	-0.37	-0.35	-0.29	-0.32	-0.52	1		
9-Num.Pubs	1169	-0.01	0.98	0.17	0.22	0.19	0.21	0.10	0.10	0.24	-0.27	1	
10-Ln(Population)	1158	7.03	0.68	-0.07	-0.08	0.00	-0.02	-0.08	-0.12	-0.01	-0.31	0.44	1
11-Fractional FWCI	1169	0.77	0.36	0.38	0.51	0.34	0.38	0.26	0.32	0.35	0.19	0.20	-0.04

Table 2 presents the cross-sectional regression models by year, from 2008 to 2015, where 2009 is missing. The coefficients are standardized to facilitate comparison between variables, and heteroskedastic consistent t-statistics are in parentheses. The samples size (N) for the models is the number of countries per year cross-section. The results show that the DemIndex is positively associated with FracFWCI in all years, except for 2011 and 2014. GlobIndex shows a strong significant correlation with FracFWCI across the entire time period, as does PercInternational and NumPubs. Ln(Population) is only significant for two years. The PercInternational and GlobIndex appear to have the largest standardized estimates. The VIF values were between two and three for democracy and globalization and under two for the remaining variables.

Table 2 – Cross-Sectional Regression by Year, DV is FracFWCI

Variables	2008	2010	2011	2012	2013	2014	2015
Dem.Index	0.328**	0.231*	0.141	0.212*	0.228**	0.116	0.180*
	(2.815)	(2.519)	(1.330)	(2.309)	(2.762)	(1.597)	(2.296)
Glob.Index	0.296*	0.347**	0.579***	0.542***	0.595***	0.430***	0.548***
	(2.426)	(3.067)	(5.158)	(5.232)	(6.238)	(4.905)	(5.836)
Perc.International	0.626***	0.527***	0.556***	0.604***	0.654***	0.545***	0.706***
	(4.979)	(4.748)	(5.145)	(6.440)	(8.862)	(5.880)	(7.225)
Num.Pubs	0.189***	0.202***	0.160**	0.173**	0.187***	0.159**	0.201***
	(3.640)	(3.779)	(3.123)	(3.166)	(3.984)	(3.173)	(4.729)
Ln(Population)	0.0277	0.0262	0.150*	0.136	0.104	0.0823	0.128*
	(0.362)	(0.373)	(2.191)	(1.854)	(1.910)	(1.402)	(2.190)
Constant	0	0	0**	0**	0***	0*	0***
	(-1.605)	(-1.188)	(-3.241)	(-2.920)	(-4.123)	(-2.264)	(-3.504)
N	161	161	161	160	160	160	160
R-Square	0.427	0.354	0.431	0.442	0.554	0.266	0.448

Standardized Coefficients, Robust t-statistics in parentheses , *** p<0.001, ** p<0.01, * p<0.05

The same models in Table 2 were run using Democ and Polity2 separately. Since the DemIndex had a strong Pearson correlation above 0.8 with both Democ and Polity2, we expected them to perform similarly against FracFWCI. However, neither variable was significant in any model when GlobIndex was included. Without GlobIndex both became positive/significant, where Democ had standardized coefficients on FracFWCI of between 0.26 and 0.385 depending on the year, and Polity2 had coefficients between 0.235 and 0.335. Both displayed a similar declining trend over time, like the DemIndex.

Table 3 shows the random effects regression results, including robust standard errors adjusted for unit(country) cluster to account for heteroscedasticity and serial correlation. DemIndex shows a positive and significant association with FracFWCI. The sub-component, political culture, shows a stronger estimate than the other measures. Political participation and functioning of government also show significant estimates on FracFWCI. GlobIndex remains statistically significant with stable estimates across the models. Unlike Table 2, these estimates are not standardized, and so estimates from different variables cannot be compared in terms of strength. The R-Square values indicate that the models account for between 35%-44% of the variance in FracFWCI with the PolCulture model accounting for the highest amount of variance; the variance explained by the random effects models is largely between unit rather than within unit. In addition, we replicated Model 1 in Table 3 using the Polity IV measures. Democ displayed a positive and significant estimate (coef=0.012, p=0.03), while Polity2 was not significant at (coef=0.006, p=0.08). Both became significant when removing GlobIndex.

Table 3 – Random Effects Regression, DV is FracFWCI (2010-2015)

Variables	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
Dem.Index	0.030**					
	(0.010)					
Pol.Culture		0.060***				
		(0.011)				
Pol.Participation			0.029**			
			(0.011)			
Funct.Government				0.020*		
				(0.009)		
Elect.Proc.Pluralism					0.003	
					(0.005)	
Civil.Liberties						0.009
						(0.007)
Glob.Index	0.009***	0.008***	0.010***	0.010***	0.011***	0.011***
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
Perc.International	0.547***	0.570***	0.567***	0.545***	0.543***	0.543***
	(0.146)	(0.139)	(0.146)	(0.146)	(0.146)	(0.147)
Num.Pubs	0.055***	0.045**	0.055***	0.054**	0.056**	0.057**
	(0.016)	(0.015)	(0.016)	(0.018)	(0.018)	(0.017)
Ln(Population)	0.011	0.025	0.007	0.007	0.006	0.009
	(0.027)	(0.025)	(0.026)	(0.027)	(0.028)	(0.027)
Constant	-0.210	-0.461	-0.221	-0.164	-0.198	-0.213
	(0.291)	(0.272)	(0.287)	(0.289)	(0.296)	(0.295)
Year	Fixed	Fixed	Fixed	Fixed	Fixed	Fixed
N	958	958	958	958	958	958
# Countries	161	161	161	161	161	161
R-Square	0.368	0.444	0.375	0.366	0.352	0.355

Robust standard errors in parentheses, *** p<0.001, ** p<0.01, * p<0.05

Finally, two-way fixed-effects regression with robust standard errors was used as a robustness check. Fixed-effects was not initially used because it ignores between unit(country) correlation and because the democracy variables have low-to-no variation for some countries in the time period. DemIndex performed less well (coef=0.035, p=0.155) than in the random effects models, while Democ (coef=0.019, p=0.023) and Polity2 (coef=0.015, p= 0.0011) performed better. However, when removing GlobIndex from these models, Democ and Polity2 became insignificant, suggesting an interaction effect between democracy, globalization, and scientific impact; this is the subject of future research.

Conclusion

This research note tests the relationships between democracy, globalization, and science. It shows strong associations between democracy and scientific impact, controlling for globalization index, population size, number of scientific papers, and percentage of papers that are internationally co-published. This note also examined the sub-components of the democracy index, where political culture appeared to have the strongest association with science. This

research also shows a strong and persistent association between globalization and science. The results warrant further research on the relationship between governmental forms, globalization, and scientific processes and outcomes.

References

Adams, G., Balfour, D. (1998). *Unmasking administrative evil*. Sage Publications Inc. Thousand Oaks, California

Chinchilla-Rodríguez, Z., Bu, Y., Robinson-García, N., Costas, R., Sugimoto, C. R. (2018). Travel bans and scientific mobility: utility of asymmetry and affinity indexes to inform science policy. *Scientometrics*, 116(1), pp. 569-590.

Durant, D. (2011). Models of democracy in social studies of science. *Social studies of science*, 41(5), pp. 691-714.

Dreher, A. (2006): Does Globalization Affect Growth? Evidence from a new Index of Globalization, *Applied Economics,* 38(10), pp.1091-1110.

Kontopoulos, K. M. (2006). *The logics of social structure* (Vol. 6). Cambridge University Press.

Kitcher, P. (2003). *Science, truth, and democracy*. Oxford University Press.

Gao, Y., Zang, L., Roth, A., Wang, P. (2017). Does democracy cause innovation? An empirical test of the popper hypothesis. Research Policy, 46(7), pp.1272-1283.

Green, E. C., & Gallery, W. (2015). "Measuring democracy". in *In Search of Democracy*, ed. Diamond, L., pp. 46-75.

Gygli, S., Haelg, F., Potrafke, N., Sturm, J.E. (2018) The KOF Globalisation Index – Revisited, *Review of International Organizations*, pp.1-32

Jasanoff, S. (2011). *Designs on nature: Science and democracy in Europe and the United States*. Princeton University Press.

Lengwiler, M. (2008). Participatory approaches in science and technology: Historical origins and current practices in critical perspective. *Science, Technology, & Human Values*, 33(2), pp.186-200.

Marshall, M. G., Jaggers, K., Gurr, T. R. (2002). Polity IV project: Dataset users' manual. College Park: University of Maryland.

Mokyr, J. (2016) *Culture of Growth: the Origins of the Modern Economy*, Princeton University Press

Nahuis, R., & Van Lente, H. (2008). Where are the politics? Perspectives on democracy and technology. *Science, Technology, & Human Values*, 33(5), pp.559-581.

Popper, K. (1966). *The open society and its enemies*. Vo.1. Princeton University Press.

Prigogine, I. (1986). Science, civilization and democracy: values, systems, structures and affinities. *Futures*, 18(4), pp.493-507.

Robinson-Garcia, N., Sugimoto, C. R., Murray, D., Yegros-Yegros, A., Larivière, V., & Costas, R. (2019). The many faces of mobility: Using bibliometric data to measure the movement of scientists. *Journal of Informetrics*, *13*(1), pp.50-63.

Simon, H. A. (1996). *The sciences of the artificial*. MIT press.

Stilgoe, J., Owen, R., Macnaghten, P. (2013). Developing a framework for responsible innovation. *Research Policy*, *42*(9), pp.1568-1580.

Stirling, A., O'Donovan, C., Ayre, B. (2018). Which Way? Who says? Why? Questions on the multiple directions of social progress. *Technology's Stories*, pp.1-20.

Taylor, M.Z. (2016). *The Politics of Innovation: Why Some Countries Are Better Than Others at Science and Technology*, Oxford University Press

Van Reybrouck, D. (2016), *Against Elections: The Case for Democracy*, Vintage Publishing

Wagner, C. S., Leydesdorff, L. (2005). Network structure, self-organization, and the growth of international collaboration in science. *Research policy*, *34*(10), pp. 1608-1618.

Wagner, C. S., Whetsell, T. A., Leydesdorff, L. (2017). Growth of international collaboration in science: revisiting six specialties. *Scientometrics*, *110*(3), pp.1633-1652

Wagner, C. S., Whetsell, T.A., Baas, J., Jonkers, K. (2018a). Openness and impact of leading scientific countries. *Frontiers in Research Metrics and Analytics*, *3*, 10.

Wagner, C. S., Whetsell, T.A., Jonkers, K., Baas, J. (2018b). Measure of National Return in International Science Cooperation. In *23rd International Conference on Science and Technology Indicators (STI 2018), September 12-14, 2018, Leiden, The Netherlands*. Centre for Science and Technology Studies (CWTS).

Waltman, L., van Eck, N.J. (2015). Field-normalized citation impact indicators and the choice of an appropriate counting method. *Journal of Informetrics, 9*(4), pp.872-894

Wiesner, K., Birdi, A., Eliassi-Rad, T., Farrell, H., Garcia, D., Lewandowsky, S., Palacios, P., Ross, D., Sornette, D. and Thébault, K. (2018). Stability of democracies: a complex systems perspective. *European Journal of Physics*, *40*(1), 014002.

Social media visibility of open access versus non-open access articles: A case study of Life Sciences & Biomedicine

Tahereh Dehdarirad[1], Fereshteh Didegah[2] and Arezoo Didegah[3]

[1] *tahereh.dehdarirad@chalmers.se*
Chalmers University of Technology, Department of Communication and Learning in Science, Göteborg, Sweden

[2] *f.didegah@ubc.ca*
iSchool, University of British Columbia, Vancouver, BC, Canada

[2] *arezoo.didegah@gmail.com*
Department of Management, Yazd University, Yazd, Iran

Abstract

This study aims to determine whether and to what extent OA status of articles predicts their social media visibility when controlling for a considerable number of important factors. These factors which previous research confirmed their positive associations with altmetric counts are journal impact, individual collaboration, research funding, number of MESH topics, international collaboration, lay summary, and F1000 Score. The data for this study comprises 83,479 articles and reviews in the research area of Life Sciences & Biomedicine from 2012-2016, retrieved from the Medline in November 2018. The results showed that (except for policy documents) the percentage of OA articles mentioned on altmetrics platforms was significantly higher than that of the NOA articles. Furthermore, OA factor is significantly associated with the increasing number of tweets, news posts, Facebook posts and blog posts and decreasing zero counts on each platform. Finally, OA status of articles is likely to increase the number of tweets, Facebooks posts, news posts, and blog posts by 97.6%, 25.2%, 78.2% and 51.4%, respectively.

Introduction

This study aims to compare the social media visibility of Open Access (OA) and Non-Open Access (NOA) articles in Life Sciences & Biomedicine. The movement towards providing open access to research outputs was initiated over sixteen years ago (Chan, et al., 2002; Piwowar, et.al, 2018) and it has been widely accepted as a desirable phenomenon and become a reality in many academic spheres (Kriegeskorte et al., 2012). The visibility of OA vs NOA articles in terms of citations has been widely investigated. Whilst some studies have suggested a causal relationship between article OA status and higher citation counts (Eysenbach, 2006; Hajjem, et al., 2006; Gargouri, et al., 2010), some other studies have identified weaknesses in the methodology used in the earlier studies where a citation advantage had been found (McCabe and Snyder, 2013; Hersh and Plume, 2016; Hua et al., 2016). OA articles have also shown a social media visibility advantage (McKiernan, et al., 2015). A study of over 2,000 articles published in Nature Communications showed that articles published openly received nearly double the number of unique tweeters and Mendeley readers than subscription-based articles (Adie, 2014). The results of a similar study on 1,761 Nature Communications articles showed that OA articles received 1.2–1.48 times as much social media attention (Twitter and Facebook) as compared to NOA articles (Wang, et al., 2015).

Some studies shed light on several factors and reasons for social media mentions of scientific outputs (Haustein et al., 2015; Andersen & Haustein, 2015; Didegah, Bowman & Holmberg, 2018). These common factors which are mainly the characteristics of articles include journal impact factor, international collaboration, individual collaboration, funding, abstract readability and abstract title (Didegah, Bowman & Holmberg, 2018). The results show that important factors vary across different social media platforms. While institution prestige and

country prestige associate with increased Mendeley readers and blog and news posts, they are insignificant factors for Twitter and Facebook posts (Didegah, Bowman & Holmberg, 2018).

Alhoori et al. (2015) compared research visibility on eight online platforms and found that OA articles received more altmetric counts than NOA articles. However, when controlling for effects of journal and publication year, no clear relationship was found between OA and altmetric mentions.

Our study follows in the same vein and aims to determine whether and to what extent OA status of articles predict their social media visibility when controlling for a considerable number of factors that might be influential. To address this, the following questions are sought to answer:

1. How are OA articles shared on different social media platforms compared to non-OA articles?
2. To what extent does the OA status of articles result in a higher social media visibility, controlling for some important factors (journal impact, individual collaboration, research funding, number of MESH topics, international collaboration, lay summary, and F1000 Score)?

Methodology

Data collection and processing

The data for this study comprised 83,479 articles and reviews in the research area of Life Sciences & Biomedicine from 2012-2016, retrieved from the Web of Science Medline in November 2018.

Using articles' PMID, a search was conducted in Altmetric.com (October 2017 version) in order to obtain the following altmetric indicators: tweet counts, Facebook posts, news posts, blog posts, F1000 post counts, F1000 score, policy mentions, and Wikipedia mentions. To answer the second research question, the number of tweets, Facebook posts, news posts, and blog posts were considered as dependent variables in a regression model (explained in the next part) to measure the extent to which the OA status of an article may predict its social media visibility. In the same model, the OA status of articles was considered as the independent variable and several other variables regarded as covariates (see Table 1). The covariates or control variables are the factors identified from the previous literature and were found to be important factors significantly influencing social media visibility of research articles (Didegah, Bowman & Holmberg, 2018). Table 1 lists all these different variables and their descriptions. We also used publication year as an offset variable in the regression model.

Data analysis and procedures

Descriptive statistics was used to depict the state-of-the-art of OA articles vs. NOA articles shared on the seven altmetrics platforms. Two-sample proportion tests were also performed in order to compare the proportion of OA papers shared on different altmetrics platforms in comparison to non-OA articles.

To answer the second research question, given that the dependent variables (altmetric counts) of this study were count data, count regression models were used. As altmetric counts are over-dispersed and include excessive number of zeros, a count model is required to deal with these two issues.

First, a standard negative binomial, a zero-inflated negative binomial and a hurdle negative binomial models were applied. A standard model is frequently used to model overdispersed data. Zero-

inflated models are used for overdispersed and excessive zero datasets and assume that there are two types of zeros in the data: zeros which arise from a negative binomial count distribution and zeros which arise from a "perfect-zero" distribution (Hilbe, 2011). Hurdle models measure the likelihood of an observation being positive or zero, and then determine the parameters of the count distribution for positive observations. We finally concluded that a negative binomial-logit hurdle model was the best fit for the data as it creates a scenario in which the positive counts follow a Poisson or NB distribution after passing a hurdle to gain positive counts (Didegah, Bowman, & Holmberg, 2018).

Table 1. Dependent variables, independent variables and covariates for the hurdle model

Variable type	Name	Measure
Dependent	Twitter	Number of tweets
	Facebook	Number of Facebook posts
	News	Number of news posts
	Blog	Number of blog posts
Independent	OA status	OA (1); non-OA articles (0)
Covariate	Individual collaboration	Number of authors collaborating in an article
	Funding	Funded article (1); not-funded article (0)
	SNIP	Source (journal) Normalized Impact per Paper
	International collaboration	International collaboration (1); national collaboration (0)
	Multi-topics	Number of MESH headings assigned to each article
	Lay summary	Articles from journals including lay summaries listed in Shailes list[1] (1); other journals (0)
	F1000 score	

Results

Question 1: How are OA articles shared on different social media platforms compared to non-OA articles?

[1]https://elifesciences.org/articles/25411?utm_source=content_alert&utm_medium=email&utm_content=fulltext&utm_campaign=elife-alerts

Out of a total of 83,479 papers, 47,957 (57.45 %) were OA and 35,522 (42.55%) were NOA. Twitter had the highest percentage of OA articles (34.56 %), followed by Facebook (8.78%), blogs (5.02%), news (4.49%), Wikipedia (1.78%), F1000 (1.76%) and policy documents (0.55%), respectively. (See Figure 1). The highest percentage of NOA articles was also shared on Twitter (15.56%), followed by Facebook (2.8%), blogs (1.33%), and news (1.29%).

The results of the two proportion tests also showed that (except for policy documents) the percentage of OA articles mentioned on altmetrics platforms was significantly higher than that of the NOA articles [P<0.0001].

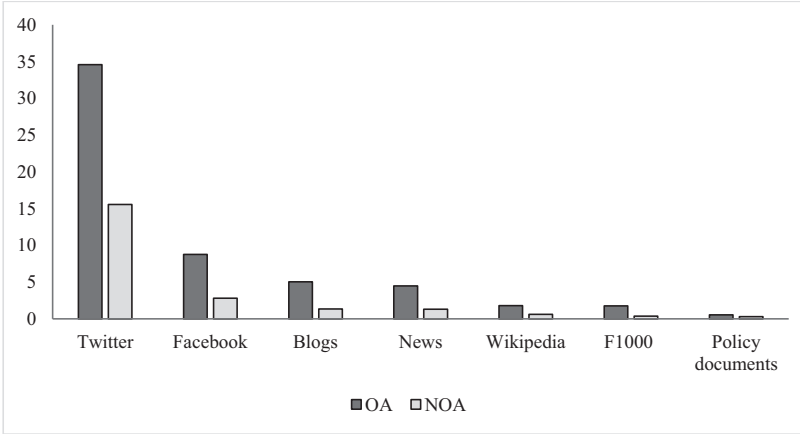


Figure 1. The percentage of OA vs NOA articles across different social media platforms

Question 2: To what extent does the OA status of articles result in a higher social media visibility, controlling for some important factors?

To measure the extent to which the OA status of an article may affect its social media visibility, a hurdle model was run investigating the association between the OA status of articles and the number of times they were mentioned on each altmetrics platform. Only Twitter, Facebook, blogs and news were considered to answer this question as there were a significant number of articles visible on these platforms (See Figure 1). To best fit the model and obtain the most reliable results, some important factors that may interfere and influence social media visibility of a single article were entered into the model as control variables.

As found from the Hurdle models (Tables 2-5), some covariates such as individual collaboration, international collaboration and the F1000 score of articles significantly associated with the increased number of tweets, Facebook posts, news posts and blog posts to articles. Some other factors such as lay summaries were important factors on few platforms but not all. However, controlling for these factors increases the likelihood of obtaining a more precise and reliable association between the OA factor and social media mentions of articles.

The hurdle model comprises two parts: the count model which is a negative binomial model and the logit model. The count model predicts the changes in the non-zero social media counts while the logit model reports the changes in the zero social media mentions for a unit change in the open access factor and each of the covariates.

Twitter

Regarding the count model, the open access factor significantly associated with the increased number of tweets and a unit change in the factor increased the number of tweets by 97.6%. The logit model confirmed that open access factor was significantly associated with the decreased number of zero tweets.

Table 2. Hurdle model results for Twitter

	Count model	Coef.	Exp.(Coef.)	Std. Error	P value[1]
	OA	0.681	1.976	0.024	<2e-16 ***
Covariates	Indiv. collaboration	0.023	1.023	0.003	<2e-16 ***
	Funding	-0.003	0.997	0.005	0.469
	No. topics	-0.064	0.938	0.002	<2e-16 ***
	SNIP	0.641	1.898	0.013	<2e-16 ***
	Int. collaboration	0.264	1.302	0.024	<2e-16 ***
	Lay summary	1.096	2.992	0.055	<2e-16 ***
	F1000 score	0.245	1.277	0.027	<2e-16 ***
	Logit model	Coef.	Exp.(Coef.)	Std. Error	P value
	OA	0.841	2.319	0.016	<2e-16 ***
Covariates	Indiv. collaboration	0.055	1.056	0.002	<2e-16 ***
	Funding	-0.013	0.987	0.004	0.0006 ***
	No. topics	-0.045	0.956	-0.001	<2e-16 ***
	SNIP	0.524	1.689	0.011	<2e-16 ***
	Int. collaboration	0.126	1.135	0.017	<7e-14 ***
	Lay summary	1.392	4.022	0.059	<2e-16 ***
	F1000 score	0.343	1.409	0.035	<2e-16 ***

[1] Sig. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Facebook

Controlling for six important factors, the open access factor significantly associated with increased Facebook post counts and it increased the number of Facebook posts by 25.2% which was the weakest association compared to the other three social media platforms. According to the logit model, the open access factor was significantly associated with decreased number of zero Facebook counts.

Table 3. Hurdle model results for Facebook

	Count model	Coef.	Exp.(Coef.)	Std. Error	P value
	OA	0.225	1.252	0.052	1.51e-05 ***
Covariates	Indiv. collaboration	0.012	1.012	0.005	0.012 *
	Funding	-0.018	0.982	0.007	0.011 *
	No. topics	-0.010	0.990	0.004	0.007 **
	SNIP	0.295	1.343	0.012	< 2e-16 ***
	Int. collaboration	0.251	1.286	0.047	6.94e-08 ***
	Lay summary	-0.087	0.917	0.082	0.29
	F1000 score	0.148	1.160	0.027	6.16e-08 ***
	Logit model	Coef.	Exp.(Coef.)	Std. Error	P value
	OA	0.833	2.299	0.027	<2e-16 ***
Covariates	Indiv. collaboration	0.027	1.028	0.002	<2e-16 ***
	Funding	0.003	1.003	0.004	0.561
	No. topics	-0.029	0.971	0.002	<2e-16 ***
	SNIP	0.442	1.556	0.009	<2e-16 ***
	Int. collaboration	0.206	1.228	0.024	<2e-16 ***
	Lay summary	0.957	2.603	0.048	<2e-16 ***
	F1000 score	0.215	1.240	0.024	<2e-16 ***

News

The results of the count model showed that open access articles likely received 78.2% more news mentions than the non-open access articles. The association was weaker than that of Twitter mentions (97.6%) but it was stronger than the Facebook mentions (25.2%). Regarding the logit model, the open access factor was significantly associated with decreased zero news mentions.

Table 4. Hurdle model results for News

	Count model	Coef.	Exp.(Coef.)	Std. Error	P value
	OA	0.578	1.782	0.078	9.03e-14 ***
Covariates	Indiv. collaboration	0.023	1.023	0.007	0.000724 ***
	Funding	0.009	1.009	0.013	0.494864
	No. topics	-0.023	0.978	0.005	6.83e-06 ***
	SNIP	0.213	1.238	0.019	< 2e-16 ***
	Int. collaboration	0.193	1.212	0.066	0.003702 **
	Lay summary	0.043	1.044	0.127	0.73322
	F1000 score	0.157	1.170	0.033	1.70e-06 ***
	Logit model	Coef.	Exp.(Coef.)	Std. Error	P value
	OA	0.965	2.626	0.039	<2e-16 ***
Covariates	Indiv. collaboration	0.039	1.039	0.003	<2e-16 ***
	Funding	-0.011	0.989	0.006	0.0585 .
	No. topics	-0.046	0.955	-0.003	<2e-16 ***
	SNIP	0.522	1.685	0.010	<2e-16 ***
	Int. collaboration	0.340	1.404	0.033	<2e-16 ***
	Lay summary	0.660	1.934	0.066	<2e-16 ***
	F1000 score	0.332	1.394	0.027	<2e-16 ***

Blogs

Open access factor significantly associated with increased blog mentions and it increased the number of blog posts by 51.4%. The logit model also showed that the open access factor was significantly associated with zero blog mentions.

Table 5. Hurdle model results for blogs

	Count model	Coef.	Exp.(Coef.)	Std. Error	P value
	OA	0.415	1.514	0.077	6.52e-08 ***
Covariates	Indiv. collaboration	0.014	1.014	0.005	0.00606 **
	Funding	-0.008	0.992	0.011	0.4534
	No. topics	-0.010	0.990	0.006	0.07417 .
	SNIP	0.250	1.284	0.017	< 2e-16 ***
	Int. collaboration	0.160	1.173	0.062	0.00982 **
	Lay summary	0.192	1.212	0.101	0.05796 .
	F1000 score	0.188	1.207	0.029	1.28e-10 ***
	Logit model	Coef.	Exp.(Coef.)	Std. Error	P value
	OA	1.046	2.848	0.039	< 2e-16 ***
Covariates	Indiv. collaboration	0.024	1.024	0.003	< 2e-16 ***
	Funding	-0.004	0.996	0.005	0.488
	No. topics	-0.045	0.956	-0.003	< 2e-16 ***
	SNIP	0.519	1.681	0.010	< 2e-16 ***
	Int. collaboration	0.259	1.296	0.032	5.06e-16 ***
	Lay summary	1.110	3.033	0.056	< 2e-16 ***
	F1000 score	0.323	1.381	0.026	< 2e-16 ***

Discussion and conclusion

Open access is certainly an important factor to help increase the visibility and impact of research articles on online platforms. The findings of our study also revealed that the percentage of Life Sciences and Biomedicine OA articles (around 57%) mentioned on social media platforms was fairly higher than the NOA articles (around 43%). However, the extent to which the OA factor may contribute to social media visibility of research outputs, by itself and regardless of any other factor interference, has not yet been clearly measured. There are numerous factors that may influence social media impact of articles and need to be controlled for their interference when measuring the actual relationship between the OA factor and number of social media mentions to the articles. This article aimed to do so through controlling for six important factors that previous research confirmed their positive associations with altmetric counts: Individual collaboration, research funding, number of MESH topics, international collaboration, lay summary, and F1000 Score. The results showed that OA factor was significantly associated with increasing number of tweets, news posts, Facebook posts and blog posts and decreasing zero counts on each platform. OA factor was strongly associated with the tweet counts and it was likely to increase the number of tweets by 97.6%. The factor was also significantly associated with a 78.2% increase in the number of news posts, 25.2% increase in the number of Facebook posts, and 51.4% increase in the number of blog posts. This was contrary to the unclear relationship that Alhoori et al. (2015) found between the OA factor and

social media impact of research articles, after controlling for journal impact factor and publication year.

The extent to which the OA factor associates with social media counts may vary by entering/removing factors to the models. But the current model attempted to control for a number of important factors with some of them being very influential. A model that controls for all influential factors may sound ideal, but it could be certainly interesting to have a more comprehensive model in a future study. This would help to more precisely predict and determine the actual association between the OA factor and social media attention of articles.

References

Adie, E. (2014). Attention! A study of open access vs non-open access articles. https://www.altmetric.com/blog/attentionoa.

Alhoori, H., Ray Choudhury, S., Kanan, T., Fox, E., Furuta, R., & Giles, C. L. (2015). *On the Relationship between Open Access and Altmetrics.* Paper presented at the iConference 2015. http://hdl.handle.net/2142/73451.

Andersen, J.P., & Haustein, S. (2015). Influence of study type on Twitter activity for medical research papers. arXiv: 1507.00154.

Chan, L., Cuplinskas, D., Eisen, M., Friend, F., Genova, Y., Guédon, J. C., ... & La Manna, M. (2002). Budapest open access initiative.

Didegah, F., Bowman, T. D., & Holmberg, K. (2018). On the differences between citations and altmetrics: An investigation of factors driving altmetrics versus citations for Finnish articles. *Journal of the Association for Information Science and Technology,* 69(6), 832-843. 10.1002/asi.23934

Gargouri, Y., Hajjem, C., Larivière, V., Gingras, Y., Carr, L., Brody, T., & Harnad, S. (2010). Self-Selected or Mandated, Open Access Increases Citation Impact for Higher Quality Research. *Plos One,* 5(10), e13636. 10.1371/journal.pone.0013636

Haustein, S., Bowman, T.D., & Costas, R. (2015). Interpreting "altmetrics": Viewing acts on social media through the lens of citation and social theories. arXiv:1502.05701.

Hersh, G., & Plume, A. (2016). Citation metrics and open access: what do we know? *Elsevier Connect.* Retrieved from https://www.elsevier.com/connect/citation-metrics-and-open-access-what-do-we-know.

Hilbe, J. M. (2011). *Negative Binomial Regression, second edition* (5th print, 2012). Cambridge, UK: Cambridge University Press.

Hua, F., Sun, H., Walsh, T., Worthington, H., & Glenny, A.-M. (2016). Open access to journal articles in dentistry: Prevalence and citation impact. *Journal of Dentistry,* 47, 41-48. doi.org/10.1016/j.jdent.2016.02.005

Kriegeskorte, N., Walther, A., & Deca, D. (2012). An emerging consensus for open evaluation: 18 visions for the future of scientific publishing. *Frontiers in computational neuroscience*, 6, 94. 10.3389/fncom.2012.00094

McCabe, M. J., & Snyder, C. M. (2015). Does Online Availability Increase Citations? Theory and Evidence from a Panel of Economics and Business Journals. *The Review of Economics and Statistics,* 97(1), 144-165. 10.1162/REST_a_00437

Erin, M., Philip E., B., C. Titus, B., Stuart, B., Amye, K., Jennifer, L., . . . Tal, Y. (2016). How open science helps researchers succeed. Retrieved from https://figshare.com/articles/The_open_research_value_proposition_How_sharing_can_help_researchers_succeed/1619902 doi:10.6084/m9.figshare.1619902.v7

Piwowar, H., Priem, J., Larivière, V., Alperin, J. P., Matthias, L., Norlander, B., . . . Haustein, S. (2018). The state of OA: a large-scale analysis of the prevalence and impact of Open Access articles. *PeerJ, 6*, e4375. doi:10.7717/peerj.4375

Wang, X., Liu, C., Mao, W., & Fang, Z. (2015). The open access advantage considering citation, article usage and social media attention. *Scientometrics,* 103(2), 555-564. 10.1007/s11192-015-1547-0

Toward Predicting Proposal Success: An Update

Caleb Smith[1], Kevin W. Boyack[2] and Richard Klavans[3]

[1] *calebs@umich.edu*
University of Michigan Medical School, Ann Arbor, MI (USA)

[2] *kboyack@mapofscience.com*
SciTech Strategies, Inc., Albuquerque, NM (USA)

[3] *rklavans@mapofscience.com*
Proposal Analytics, Inc., Wayne, PA (USA)

Abstract

Indicators that could predict the success or failure of 3459 research proposals are identified and evaluated. The sample was highly homogeneous (all proposals were from one medical school and submitted to one funding agency) but heterogeneous within this context (all types of NIH proposals are included). The most important exogenous indicator was whether the PI had a backlog of proposal opportunities. Gender and race had no statistically significant impact. Only one of the six linguistic indicators (derived from the research strategy section of these proposals) was a significant predictor of proposal success.

Introduction

Research proposals are unique documents. Principal investigators (PI) spend a great deal of time and effort convincing potential funders that they can help solve the specific research problems that the funder is interested in. Proposals typically precede the outputs that researchers are evaluated on. Proposals precede grants. Proposals precede publications that report findings from funded research. Proposals precede patents – there is usually a proposal (a structured request for money to 'do the work') that occurs before the patentable invention. The content in a proposal provides one of the earliest signals of a potential breakthrough.

Funding agencies have taken an active lead in analyzing the reasons that they fund (or don't fund) a research proposal. Most of this literature addresses the worldwide concerns of diversity, gender bias and innovativeness. Funding agencies within the United States also look at possible racial biases in proposal evaluation vis-à-vis Black/African Americans, Native Americans, Asians and Hispanics.

But the research institutions that submit these grants have been relatively silent on this issue. This is somewhat surprising because it is the research institutions that hire and promote faculty, and presumably provide assistance to their research faculty so that they are more effective and efficient. And it is curious to us that the peer reviewed literature is rather silent on these issues from the point of view of research institutions.

The University of Michigan has initiated what may be the first large scale analysis of grant proposal success and failure from the perspective of a submitting organization. This project builds upon a 'proof of principle' study (Boyack, Smith, & Klavans, 2018) that suggested that the text in the document is more likely to yield indicators that predict proposal success than traditional bibliometric indicators. This study expands on the earlier study by developing and testing multiple text-based indicators.

The structure of this study is as follows. In the literature review, we point out that prior research tends to focus on biases associated with gender, race and innovativeness from the perspective of the funding agency. Prior research does not look at this phenomenon from the

perspective of the submitting organization or the submitting researcher. The literature review is followed by a description of our hypotheses and database. Exogenous indicators (outside the control of the researcher) and endogenous indicators (within the control of the researcher) are identified. Multi-stage analysis is then used to estimate the impact of exogenous factors before determining the marginal impact of the endogenous factors. Overall, only one of the five proposed endogenous indicators had significant predictive value. We conclude by discussing implications of this work, along with limitations and potential future directions.

Prior literature

The literature on proposal success/failure is presented from two perspectives. Most of the literature looks at this issue from the perspective of funders and the corresponding need to create fair and effective mechanisms for funding research. The second perspective is from the perspective of the submitting institution or researcher and the corresponding need is to make individual researchers more efficient and effective.

National science policy

The largest-scale empirical studies of proposal success in the US are those performed by Ginther and colleagues (Ginther et al., 2018; Ginther, Haak, Schaffer, & Kington, 2012; Ginther, Kahn, & Schaffer, 2016; Ginther et al., 2011). These studies were based on 83,000 R01 proposals submitted to the NIH between 2000 and 2008. Ginther combined proposals that were related (new proposals are combined with subsequent resubmittals and then treated as a single observation).

Ginther focused on the prior training, experiences and publication record of the PI as an exogenous variable. In the short term, PIs have no control over their cumulative 'academic wealth' when they submit a proposal. Other exogenous variables were considered such as the rank of the submitting institution, the PI gender and PI race. No endogenous variables were considered. There was no analysis of the text or references in the proposals, and there was no consideration of the amount of money that was being requested. The purpose of this stream of research was to determine if there was evidence of racial and gender bias. While there was strong evidence that such differences existed, the interpretation of the causes dealt mostly with the possibility that, in the United States, white males tended to have accumulated more academic wealth at each stage of their career paths. This interpretation is consistent with related studies that show that gender and racial bias in research can be traced back to early educational experiences and social biases that result in structural inequity.

A recent empirical study of 1742 European Research Council life science proposals did find evidence of gender bias in some parts of the field after controlling for facets of past performance, and that review panel characteristics may have a role in that bias (Van den Besselaar, Schiffbaenker, Sandström, & Mom, 2018). Using roughly the same data, Mom et al. (2018) showed that probability of success also increases if a near-by panelist is from the same institution as the applicant.

Although we are unaware of studies that look at a potential bias against innovative proposals, a recent study of 18,476 research proposals submitted to the Australian Research Council's Discovery Program showed a strong bias against interdisciplinary proposals (Bromham, Dinnage, & Hua, 2016). Conformance is a common criticism aimed at funding agencies (Nicholson & Ioannidis, 2012). The biggest impediment to progress on this issue, however, may be that there are no validated indicators of proposal innovativeness. We have yet to tease out the difference between intended innovativeness (which would be signalled in the

proposal) vs. realized innovativeness (which would be signalled by the stream of publications and inventions that might result from the grant).

Increasing the efficiency and effectiveness of researchers

It takes a significant amount of time to write a research proposal. Two recent studies observed an average time commitment of 21 working days for proposals in astronomy and psychology (Von Hippel & Von Hippel, 2015) and 38 working days (Herbert, Barnett, Clarke, & Graves, 2013) for new proposals in medicine. While these numbers are not directly comparable due to differences in country, agency, and proposal scope, they both indicate a sizeable time investment, and it is reasonable to assume that a typical proposal for a sizeable research project will require a minimum of a month to write. Herbert et al. (2013) also found that more time spent on a proposal did not increase its success rate, and that resubmitted proposals did worse than new proposals. We know very little about the relationship between time spent on a proposal and the amount awarded. But it is common knowledge that, in the United States, the size and nature of these grants have a direct effect on the PI's likelihood of promotion and tenure, especially if the grant 'travels with the PI' (i.e., the money goes with the PI if the PI decides to move to a different university).

Overall, although grant money is essential to the academic research enterprise, we know relatively little about what a successful proposal consists of. This is particularly true for those factors that are within the control of the individuals writing their proposal. There are dozens of books and articles on 'how to write a successful research proposal'. However, none of these documents are rooted in statistical studies that have been subjected to rigorous peer review. One could reasonably say that our knowledge about this issue is more folklore than fact.

Commonly held beliefs about 'what makes for proposal success' have no empirical basis. As examples, some believe that clarity of writing is important, but there are no studies that use established metrics of writing clarity to predict proposal success. There is recent research that suggests a difference in the language that men and women use in research (Thelwall, Bailey, Tobin, & Bradshaw, 2019), but there has been no investigation into the possibility that 'male' vs. 'female' language has any effect on proposal success. PIs 'tell a story' in the research strategy section of a proposal. However, we don't know what types of stories are told and whether one type is more likely than another to increase proposal success.

We have addressed some of these issues in a relatively small (369 R01 proposals) 'proof of principle' study conducted for the University of Michigan (Boyack et al., 2018). Our research questions focused on commonly held beliefs about the following. We posited that proposal success would be based on writing quality; institutional reputation (on the specific research topic associated with proposal); PI reputation (what topics the PI has published) and the match between the proposal and PI topics. What is particularly surprising about this study is the lack of evidence that most of these factors made a difference in proposal success. Writing quality provided the strongest signal of predictive value. The following study is a follow up on the hypothesis that writing makes a difference.

Data

The research proposals used in this study represented all research proposals submitted to NIH by the University of Michigan that utilized the Public Health Service application package (PHS 398) between 2010 and 2016. Of the 3793 proposals in this database, only 19.6% were funded. This is slightly higher than the overall acceptance rate of 18.2% for the full set of 353,952 NIH grants over the same time period (source: https://report.nih.gov/nihdatabook/).

Table 1 shows the temporal pattern of proposals, proposal acceptance rate and submittal rates by Women or Black African/Americans PIs by year. The sudden increase in proposals from 2010 to 2011 is not due to an increase in productivity. Rather, the sample is based on an increase in the use of the PHS 398 application package. During this seven-year period, the percentage of Women and Black PIs remain relatively constant. Both percentages are slightly above the averages reported for the general population of PIs submitting proposals to NIH.

Table 1. Gender and race characteristics for principal investigators.

Year	*# Proposals*	*% Funded*	*% Women*	*%Black*
2010	353	20.4	29.5	1.42
2011	499	17.4	31.1	2.61
2012	571	18.4	31.0	1.58
2013	566	18.0	31.1	0.89
2014	591	21.0	32.8	1.86
2015	619	21.3	30.4	1.62
2016	594	20.4	34.3	1.85
All	3793	19.6	31.6	1.69
NIH Baseline		18.2[1]	29.9	1.61[2]

[1] source: https://report.nih.gov/nihdatabook/
[2] source: Ginther et al. [2011] (covers 2006-2010 time period and only for R01 grants)

3459 proposals (91.2%) were submitted by the Medical School. Of these proposals, 62% were for R01 grants (the most common type of grant awarded by NIH). The next most common grant types were R21 (N=887) and R03 (N=150). Overall, 95.3% of the proposals were R-type grants intended for researchers. 82.7% of the sample had only one PI, and when there were multiple PIs, we selected the investigator identified by the study team as the primary point of contact for the project. There were an additional 334 proposals that were submitted to NIH from the College of Engineering. We are still in the process of collecting relevant data on the proposals from the College of Engineering and they are correspondingly excluded in our preliminary analysis.

Information about the sample of proposals by PI rank (from instructor to full professor) is presented in Table 2. Of the proposals submitted to the Medical School, only 40.4% were submitted by PIs with a medical degree. The percent of PIs with medical degrees increases dramatically with PI rank. Only 11.2% of the instructors who submitted proposals had an MD degree while 48.9% of the full professors had an MD degree.

Table 2. Rank characteristics for principal investigators.

Year	*# Proposals*	*% MD*	*% Women*	*%Black*
Instructor	169	11.2	5.2	2.0
Asst Professor	1078	33.9	35.0	43.8
Assoc Professor	880	41.0	31.8	22.9
Full Professor	1332	48.9	27.9	31.2

There were only 902 PIs that submitted proposals from the medical school between 2010 and 2016. Of these, 249 (27.6%) only submitted one proposal. The median number of proposals was 3 and the average was 3.8 (one PI submitted 35 proposals during this seven-year period).

Of the 614 PIs who were not full professors, 138 were promoted (2011 through 2016). Neither race nor sex had any statistical impact on the likelihood of promotion.

Hypotheses

Our hypotheses are framed in terms of five exogenous factors and five endogenous factors. Each is assumed to have a strong influence on the likelihood of proposal success. Exogenous factors – those that cannot be controlled by the PI – are listed first.

Backlog: Whether a PI has a backlog of prior proposals is based on the NIH classification that a proposal is a resubmittal. Resubmittals are de facto evidence of a backlog. We therefore coded a proposal as a '0' if it was considered a new proposal and a '1' if it was a resubmittal. This variable should be significantly related to proposal success rates.

Gender: Consistent with the assumption that gender bias favors men, female PIs were coded as a zero and male PIs were coded as a 1. A positive relationship with proposal success would be evidence of gender bias toward males.

MD degree: Ginther found that PIs with a medical degree have higher proposal success rates. We therefore developed a dummy variable to capture this potential effect.

Race: Racial classifications of PIs are extremely problematic. In Ginther's studies, only 56% of the PIs could be placed into the four racial categories she selected. In contrast, almost all of the PIs in this sample (96.5%) could be placed into these same four categories. The remaining 34 individuals in our sample (who self-reported as being mixed or were unwilling to comply with the request for racial identity) were placed into the four racial categories by the first author based on information readily available on the internet. Proposal success rates reported by Ginther et al. (2011) for these four racial categories are 16.6% (Black), 25.5% (Asian), 28.1% (Hispanic), and 29.4% (White). We have thus constructed a variable ranging from 16.6 to 29.4 for race based upon this ordering of success rates.

The difference between success rates for Blacks and Whites is consistent with commonly held assumptions about racial bias. More interesting, perhaps, is the ordering of Asian and Hispanics within this framework. This ordering is consistent with Linneaus' premise that the continental location of ancestry matters. The ordering is from the most eastern location on the Asian continent to the most western location. The British Isles (the primary location of the founding founders of the United States) is on the western edge of the Asian continent. Hispanics (people whose ancestors emigrated from Spain and Portugal) are closer to the extreme western edge of Asia than India or China (central and eastern Asia). We posit that racial bias continues to have a significant effect on proposal success.

Salary: Rather than using an ordinal variable to represent professional rank, we use average salary for each rank (see Table 3) in order to highlight the degree of academic wealth. With a rise up the professorial rank also come increased experience, increased numbers of papers and citations, and increased access to both human and material resources. The increase in experience and resource may make it easier for the full professor to write many proposals whereas lack of experience and access to resources may make it more difficult for the instructor to write a proposal. We expect salary (as a proxy for many things) to be a significant factor in increasing backlog (generating a potential problem of collinearity in our attempt to predict proposal acceptance rates).

Table 3. Average salary as a function of professorial rank (source: umsalary.info).

Instructor	*Asst Professor*	*Assoc Professor*	*Full Professor*
64,825	110,892	125,968	183,811

We now focus on the endogenous factors – those that can be controlled by the PI.

Program type: A PI can choose which type of grant to submit to NIH. Ginther focused only on R01s (by far the largest category). This study considers all possible types. Information on the size of the proposals (both in terms of the number of words and the amount of money that is asked for) will be analysed to determine how to best capture this effect.

Proposal size: The amount of text in the research strategy sections of the NIH proposals varies widely. Given the highly skewed distribution, we use a log transform of the amount of text as an indicator of proposal size.

Grant size: Here we use the total direct costs listed in the budget section of the grant. Since this is a highly skewed variable, we used a log transform as the indicator of grant size.

Our working hypothesis is that proposal size (the number of words in the research strategy section) is associated with grant size (how much of your time will be paid for in the grant). One works harder to get larger grants. We expect that larger grants will be more competitive and correspondingly have a lower proposal acceptance rate.

Gunning fog index: One of the normative claims on 'how to write a research proposal' is that it should have a high level of clarity. We use the Gunning fog index – a simplistic measure of 'hard to read' text – as a measure of clarity of the research strategy section of each proposal. The fog index determines the assumed educational level of the reader (elementary, junior high, high school, college or PhD level).

Male-oriented language: Thelwall et al. (2019) recently identified the differential word choices in scientific articles by male and female authors. His list of the top 100 male and female words were used to create an indicator of male-oriented language. The equation we used was log(#male words+1)-log(#female words+1). The working hypothesis is that male-oriented language has a positive effect on proposal success.

Positive emotional language: A PI can choose to use words that signal positive or negative emotion. We developed an indicator of positive vs. negative emotional content. Emotional values were assigned based on the NRC Emotion Lexicon (Mohammed & Turney, 2013). The formula for the indicator was log(#positive words +1) – log(negative words+1). This is an exploratory variable – we have no *a priori* belief that positive vs. negative language has an effect.

Story arc: While the overall level of emotional language may or may not make a difference to proposal success, the temporal pattern of emotional language has been used to characterize different story arcs (Reagan, Mitchell, Kiley, Danforth, & Dodds, 2016). Many fiction stories (such as Cinderella) will start out positive, change to a long period of negativity and then finish on a positive note. For this study we used the Syuzhet Sentiment Dictionary developed by the Nebraska Literary Lab, which was chosen for its lexical scope and ability to generate

meaningful emotional arcs. The extracted research strategies were each segmented at the sentence level. Each word was then assigned a value between -1 and +1 according to its emotional valence. These values were then summed at the sentence level. Given the imprecise nature of communication and emotion, the raw data generated in this way is likely over precise. In addition, other authors have found that dictionary-based methods of sentiment analysis perform poorly at the individual sentence level (Reagan, Danforth, Tivnan, Williams, & Dodds, 2017; Ribeiro, Araujo, Goncalves, Goncalves, & Benevenuto, 2016). We avoid the risk of precision bias by smoothing and normalizing the raw values using inverse discrete cosine transformation and low pass filtering. We return 100 values for each text via decimation to facilitate further comparison.

In order to group emotional arcs by shape, we use the Time-series Anytime Density Peaks (TADPole) clustering method as proposed by Begum, et. al. (2015), which is a framework adapted to time-series similarity clustering through Dynamic Time Warping (DTW). DTW seeks to find an optimal match between temporal emotional patterns that may vary in speed by compressing or expanding sections of the time axis for one vector or the other thus allowing greater flexibility in shape-matching than would be offered by other measures such as Euclidean distance. The optimal clustering solution was identified by calculating the Calinski-Harabasz index (Arbelaitz, Gurrutxaga, Muguerza, Perez, & Perona, 2013). Eleven cluster solutions were calculated with the six-cluster solution clearly scoring the highest. This is an exploratory variable. We are testing the assumption that there is 'one best way' to tell a story in the research strategy section of a research proposal.

Results

Stage 1: Probit analysis was used in the first stage of analysis. This stage only involved the exogenous variables. The results of this analysis are shown in Table 4. By far, the single most important factor is whether there is a backlog of (previously unfunded) proposals for the PI to work on. The Z statistics for the remaining variables are relatively low (none are significant at the .001 level). The effect of gender and race are clearly not significant. There was no evidence of collinearity – the largest Pearson correlation between the independent variables was 0.155).

Table 4. Relationship between proposal success and exogenous variables.

```
Probit regression                                 Number of obs   =       3459
                                                  LR chi2(5)      =     275.03
                                                  Prob > chi2     =     0.0000
Log likelihood = -1579.7057                       Pseudo R2       =     0.0801

------------------------------------------------------------------------------
      funded |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     backlog |   .7776319    .051694    15.04   0.000     .6763136    .8789502
      salary |   2.30e-06   7.35e-07     3.13   0.002     8.61e-07    3.74e-06
          md |   .1217842   .0519294     2.35   0.019     .0200043     .223564
        male |   .0677719   .0554888     1.22   0.222    -.0409842    .1765279
        race |   .2552627   .3499462     0.73   0.466    -.4306192    .9411445
       _cons |  -1.828421   .3348066    -5.46   0.000    -2.484629   -1.172212
------------------------------------------------------------------------------
```

Stage 2: The probit model was used to estimate residual probabilities of acceptance (an interval variable that represents unexplained variance in the data). Residual probabilities are correspondingly used in Stage 2 to determine if endogenous (controllable) factors can improve our ability to predict proposal success.

Our first task, however, was to look for collinearity and adjust our analyses accordingly. As expected, the correlation between proposal size (number of words) and grant size (direct costs) was highly correlated (0.795). We therefore tested whether there was a general relationship between proposal size and grant size across any program type with at least 10 observations (see Figure 1). Given the relationship between grant size and the number of words in the research strategy, we decided to test both approaches independently.

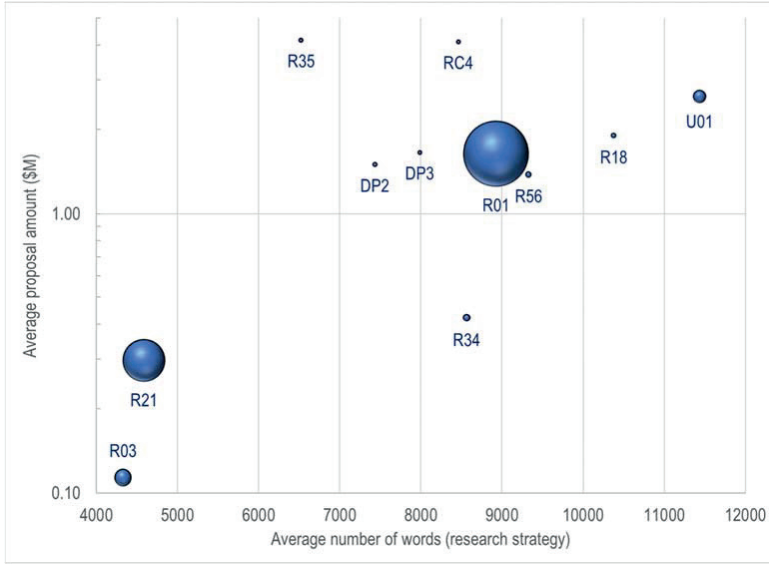


Figure 1. Relationship between proposal size and amount by grant type. Circle sizes reflect the number of proposals of each type.

The second issue we needed to resolve was to determine if there was a Story arc that was 'best'. Table 5 summarizes the average residual probabilities for each of the six Story arcs. Story arc #6, which represented 38% of the proposals in this sample, was clearly the best (they had the highest value for proposal acceptance rates both before and after adjusting for exogenous factors). We therefore generated a dummy variable for story arc #6 to determine if this Story arc has a significant residual effect on proposal success.

Table 5. Probabilities of acceptance for six different story arcs.

Story Shape	*# Proposals*	*Average. Probability*	*Avg. Residual Probability*
1	157	14.6%	-5.04%
2	137	19.0%	-1.19%
3	355	17.5%	-3.16%
4	344	20.6%	0.20%
5	1326	17.6%	-1.21%
6 (best)	1140	23.4%	3.17%

ALL	3459	19.7%	0

At this point, we did a regression analysis where the dependent value was residual proposal acceptance and the independent values were size (either log(number of words) or log(amount of the grant)), story arc (0,1 variable for the best Story arc), the fog index, the use of male language or whether there was an overall positive emotional content in the text in the research strategy section of the proposal.

Table 6 shows that the only indicator that explains residual values in proposal acceptance rates (at a p<.01 level) is the Story arc. Size (the number of words or the amount of money) may have a small role, but the results are not significant. Nor does the educational level of the written description (high school to PhD level) seem to matter. Since the Story arc is based on the pattern of emotional words over time, it is interesting to see that a simple indicator of overall type of emotion had no explanatory value in this sample.

Table 6. Summary of regression coefficients for predicting residual proposal acceptance.

	Using proposal size		*Using grant size*	
	T	*P > |t|*	*T*	*P > |t|*
Proposal Size	2.16	.031	-	-
Grant Size	-	-	-1.59	.112
Best Story	3.26	.001	3.46	.001
Fog Index	-1.15	.250	-1.17	.244
Positive Emotion	.54	.590	.35	.724
Male Language	-.52	.604	-.73	.468

Discussion and implications

We have often used a metaphor of the research system where we think of researchers as mice that self-organize around cheese (i.e., funding). Researchers receive grants and then produce papers, and the resulting piles of papers are, in essence, the consequence of where the cheese has been placed. This logic has driven our prior work on modeling the structure of research using bibliometrics and, more recently, predicting (with exceptionally high levels of accuracy) the amount of funding associated with each pile of papers (Klavans & Boyack, 2017).

We also realize that those piles of papers represent only a small fraction of the research that has been proposed. Roughly 80% of proposed research is not funded. How should we model failed proposals? We believe that researchers should rely on the insights from failed proposals and this belief comes from observations on both sides of the fence. Having served on review panels for funding agencies, we have seen the potential value of proposals with good ideas that are not presented to best advantage, and have given feedback aimed at helping the researcher to improve the proposal. As submitters of proposals that have not been funded, we greatly appreciate constructive feedback from those who have reviewed our proposals. This personal connection affects how one responds to failure.

An understanding of how science evolves as a response to failure is something that we don't consider often enough. We know how researchers respond to a negative review of a paper. Some will take the advice of reviewers, make substantial changes and additions to their work, and will ultimately get the paper accepted in the original venue. Others will quickly submit the paper to a journal of lower impact. Others will abandon the paper entirely. We suppose

that researcher responses to negative reviews of proposals are very similar, but we have no access to data that would shed light on this question.

One of the motivations of this study has been to provide a concrete way to address racial and gender bias in the funding of research in the United States. One interpretation of the results of this study is that for those just beginning an academic career, priority could be put on building up an inventory of high-priced (but not necessarily well-written) proposals … that are almost guaranteed to fail.

This may seem to be a strange interpretation of the data but consider the following. Writing quality doesn't seem to matter very much. The dollar size of the grant doesn't matter much, so asking for more money isn't going to reduce the probability of success. Having ideas that have been previously vetted (i.e., backlog) does seem to matter. This concept, while perhaps new in the context of research proposals, is not a novel one. Within the arts, it is relatively common to cultivate such backlogs of creative output. Notably, these artists would not often consider such backlogged works 'failed' but rather 'unfinished.' Perhaps they will be refined at some future point based on reviewer/editor/curator feedback, perhaps they will be 'workshopped' among peers, or perhaps this feedback will be used to create something completely novel. Perhaps they will simply be resubmitted when the cultural zeitgeist is more obviously welcoming. (How likely might a proposal to study the Zika virus have been prior to 2015?)

Due to the difficulty in obtaining data about unfunded proposals, administrators and faculty are often exposed to only small numbers of failures and very large numbers of successes. It is therefore likely that the advice we give to our students, fellows, and faculty suffers heavily from survival bias. When we make generalizations about proposals, we make generalizations about successful proposals. At an institutional level, we cannot yet make defensible claims regarding the differences between success and failure, and without this information we cannot properly educate our faculty or intervene systematically on their behalf. Proposal writing has been treated by academic institutions as more art than science, with all of the limitations implied by that statement, but few of the positives. Rather than incorporating the best practices and lessons learned from other creativity driven fields (and few scientists indeed would argue that science is not ultimately a creative endeavour), unsuccessful proposals are left unexamined in any meaningful way.

Academic institutions expend significant effort in identifying their research strengths. Additional value could be found in identifying institutional research weaknesses. In what topics do faculty proposals fail at uncommonly high rates? These are areas ripe for institutional intervention. What proposal writing strategies can we show are more likely to result in higher rates of success? These techniques should be taught and encouraged. Can we detect biases in the success of certain demographics in specific areas? These are lessons that our faculty leadership and sponsor agencies should be made aware of.

Our preliminary data show that failure is not final but can rather be used as a springboard for future success. Building up a backlog (or inventory) of ideas means that a researcher has spent their time trying to communicate dreams by writing proposals. The fact that those dreams were not funded suggests improvement is needed. A natural extension of our results is that feedback from reviewers can be a guide to where future efforts can be spent. It also suggests that this feedback could be analysed by institutions to differentiate between successful and

unsuccessful proposal tactics. We have the opportunity to learn how to productively reinterpret the failure of individual submissions in a way that encourages future success.

Limitations and future directions

This is a preliminary report and has a number of limitations. We need additional knowledge of what has been done by others to contextualize our results. Our data requires further development. We still have missing data from the proposals that were submitted from the College of Engineering. We continue to find anomalies in the data (such as two proposals where there doesn't seem to be any request for money) that require going back and rechecking the data. Converting pdf files and extracting relevant information is a very messy business.

Additional indicators that may pick up on other aspects of 'what makes for a good research strategy section' need to be developed. Although the Story arc indicator looks to be very promising, much more work is needed to identify textual features that could be good predictors of proposal success.

It is also important to point out that this study has not yet accounted for bibliographic information that is submitted with the description of the research strategy or the bibliographic information that is associated with the PI. Given that the most important predictor appears to be 'backlog', we have the possibility of generating more refined indicators of this phenomena. Using the bibliographic profiles of proposals, we can accurately determine the degree to which one proposal builds upon a prior proposal. We can tell if a proposal builds on the PI's expertise. And by using these bibliographic profiles, we hope to be able to develop an indicator of innovativeness that will be predictive in nature.

Acknowledgments

We acknowledge the Deans of the University of Michigan Medical School and College of Engineering for seeing the value of analyzing proposals and overcoming fears and trepidations associated with making some of this information available for exploratory analysis.

References

Arbelaitz, O., Gurrutxaga, I., Muguerza, J., Perez, J. M., & Perona, I. (2013). An extensive comparative study of cluster validity indices. *Pattern Recognition, 46*, 243-256.

Begum, N., Ulanova, L., Wang, J., & Keogh, E. (2015). *Accelerating dynamic time warping clustering with a novel admissible pruning strategy*. Paper presented at the Conference on Knowledge Discovery and Data Mining (KDD '15).

Boyack, K. W., Smith, C., & Klavans, R. (2018). Toward predicting research proposal success. *Scientometrics, 114*(2), 449-461.

Bromham, L., Dinnage, R., & Hua, X. (2016). Interdisciplinary research has consistently lower funding success. *Nature, 534*, 684-687.

Ginther, D. K., Basner, J., Jensen, U., Schnell, J., Kington, R., & Schaffer, W. T. (2018). Publications as predictors of racial and ethnic differences in NIH awards. *PLoS ONE, 13*(11), e205929.

Ginther, D. K., Haak, L. L., Schaffer, W. T., & Kington, R. (2012). Are race, ethnicity, and medical school affiliation associated with NIH R01 type award probability for physician investigators? *Academic Medicine, 87*(11), 1516-1524.

Ginther, D. K., Kahn, S., & Schaffer, W. T. (2016). Gender, race/ethnicity, and National Institutes of Health R01 research awards: Is there evidence of a double bind for women of color? *Academic Medicine, 91*(8), 1098-1107.

Ginther, D. K., Schaffer, W. T., Schnell, J., Masimore, B., Liu, F., Haak, L. L., et al. (2011). Race, ethnicity, and NIH research awards. *Science, 333*, 1015-1019.

Herbert, D. L., Barnett, A. G., Clarke, P., & Graves, N. (2013). On the time spent preparing grant proposals: An observational study of Australian researchers. *BMJ Open, 3*, e002800.
Klavans, R., & Boyack, K. W. (2017). Research portfolio analysis and topic prominence. *Journal of Informetrics, 11*(4), 1158-1174.
Mohammed, S., & Turney, P. (2013). Crowdsourcing a word-emotion association lexicon. *Computational Intelligence, 29*(3), 436-465.
Mom, C., Sandström, U., & Van den Besselaar, P. (2018). *Does cronyism affect grant application success? The role of organizational proximity*. Paper presented at the 23rd International Conference on Science and Technology Indicators (STI 2018).
Nicholson, J. M., & Ioannidis, J. P. A. (2012). Conform and be funded. *Nature, 492*(7427), 34-36.
Reagan, A. J., Danforth, C. M., Tivnan, B., Williams, J. R., & Dodds, P. S. (2017). Sentiment analysis methods for understanding large-scale texts: A case for using continuum-scored words and word shift graphs. *EPJ Data Science, 6*, 28.
Reagan, A. J., Mitchell, L., Kiley, D., Danforth, C. M., & Dodds, P. S. (2016). The emotional arcs of stories are dominated by six basic shapes. *EPJ Data Science, 5*(31).
Ribeiro, F. N., Araujo, M., Goncalves, P., Goncalves, M. A., & Benevenuto, F. (2016). SentiBench - A benchmark comparison of state-of-the-art practice sentiment analysis methods. *EPJ Data Science, 5*, 23.
Thelwall, M., Bailey, C., Tobin, C., & Bradshaw, N.-A. (2019). Gender differences in research areas, methods and topics: Can people and thing orientations explain the results? *Journal of Informetrics, 13*(1), 149-169.
Van den Besselaar, P., Schiffbaenker, H., Sandström, U., & Mom, C. (2018). *Explaining gender bias in ERC grant selection - Life sciences case*. Paper presented at the 23rd International Conference on Science and Technology Indicators (STI 2018).
Von Hippel, T., & Von Hippel, C. (2015). To apply or not to apply: a survey analysis of grant writing costs and benefits. *PLoS ONE, 10*(3), e0118494.

An exploration of the concept of complementarity over knowledge spaces in firm acquisitions

Lu Huang[1], Qiuju Zhou[2], Chang Wang[3], Jos Winnink[4], and Ismael Ràfols[5]

[1]*huanglu628@163.com*
School of Management and Economics, Beijing Institute of Technology, Beijing, 100081, P.R. China

[2]*zhouqiuju@casisd.cn*
Institutes of Science and Development, Chinese Academy of Sciences, Beijing, 100190, P.R. China

[3]*wangchang9611@163.com*
School of Management and Economics, Beijing Institute of Technology, Beijing, 100081, P.R. China

[4]*winninkjj@cwts.leidenuniv.nl*
Centre for Science and Technology Studies (CWTS), Leiden University, Leiden, The Netherlands

[5]*i.rafols@ingenio.upv.es*
Ingenio (CSIC-UPV), Universitat Politècnica de València, València, Spain,
Centre for Science and Technology Studies (CWTS), Leiden University, Leiden, The Netherlands

Abstract

This paper proposes a new operationalisation of the concept of scientific and technological complementarity, using firm acquisitions as case studies. In many previous studies, the probability of collaboration or acquisition is based on the degree of cognitive similarity between two laboratories or firms. Here, we propose that the scientific or technological complementarity of the acquired firm should be taken into account. We define complementarity as the non-overlapping expertise within the scope relevant for the collaboration or acquisition. For the operationalisation of complementarity, we highlight that the underlying knowledge space needs to be considered, as shown in maps of science or patents. We illustrate this operationalisation with a sample of 24 case studies of technology-driven M&As in the Chinese healthcare industry.

Introduction

Technology-driven mergers and acquisitions (Tech M&As) enable firms to have access to new knowledge and capabilities, e.g., research and development (R&D) skills, experienced personnel, and specific new technologies (Amesse et al., 2004; Williamson, 1983). Thus, Tech M&As work as an external source of innovation (Heeley et al., 2006). For strategic decision making, knowledge about current scientific and technological research activities within the working area of a company is indispensable. The analysis of the evolutionary relationship of science and technology knowledge for decision support cannot be separated from the essence of knowledge evolution. The revealing of the internal mechanism of knowledge transfer needs to deepen the relationship between knowledge units and their mutual relations within science and technology. Some researchers elucidate how the combination of technological relatedness and product relatedness between acquiring and target firms affects post-innovation performance of technology-driven M&As (Zhou et al., 2018). Existing literature indicates that technology relatedness analysis is relevant for selecting firms in Tech M&As (Ma et al., 2017). Technological relatedness involves two aspects: technological similarity and technological complementarity (Makri and Lane, 2010). Existing investigations on technological relatedness are mostly conducted from the perspective of technological similarity and some significant observations were identified. For example, technological similarity would negatively affect R&D productivity and result in research reduction (Hitt et al., 1991; Paruchuri and Eisenman, 2012; Ranft and Lord, 2002). Of the existing investigations on technological relatedness only a few are conducted from the

perspective of technological complementarity. Existing studies highlight that complementary technological resources between two firms could play positive roles in enriching the acquiring firm's knowledge base and offering opportunities for inventions (Cassiman et al., 2005; Wulf and Singh, 2013). The acquisition of firms with technological assets related to those of the acquiring firm leads to better performance than acquisitions with unrelated assets (Hussinger, 2010). It was also found that maximum performance can be realized when the two firms' technology portfolios are related (Cantwell et al., 2004), but not too similar.

The concept of complementarity

Existing studies mostly employed the definition given by Makri and Lane (2010). This defines technology complementarity between companies as the degree to which their technological problem-solving ability focuses on different narrow-defined areas of knowledge within a broad-defined area of knowledge that they share. In real-world Tech M&As, the transaction must be guided with specific purposes, e.g., the intention to develop an acquirer firm's technology ability in a particular technology area (James et al., 1998). In this vein, we consider technological complementarity as the degree that an acquired firm could supplement or add to an acquiring firm with complementary technological capabilities, in a defined technological area serving to the technological motivation of M&As. More precisely, our study focuses on horizontal acquisitions (Ma et al., 2017), in which the R&D operations of acquiring and acquired firms generally are in the same broad-defined technological area but with different degrees of overlap, ensuring the relevance of the acquisition.

There are two key factors we need to pay attention to when describing technological complementarity in M&As: 1) the technological scope U; 2) the relevant technology areas of the acquiring company A and the acquired company B. An example is given in Fig. 1. We denote U and U' as two different technological scopes. Firm A is an acquiring firm, and firms B and C as the potential acquired firms. It is apparent that A and C have no complementarity because C is outside the relevant scope U. However, there would be complementarity for scope U'. Therefore, it is necessary to define a specific technological scope in advance when investigating technological complementarity. One of the main motivations of an M&As event is the technological scope (James et al., 1998).

The relevant technology areas of two firms can be different. For example, the technology area of firm A is bigger than firm C within scope U' in Fig. 1, and A contributes more to the technological complementarity of C than the inverse. In the context of M&As, we select the acquiring firm as the main subject, and focus on the technological complementarity of the acquired firm.

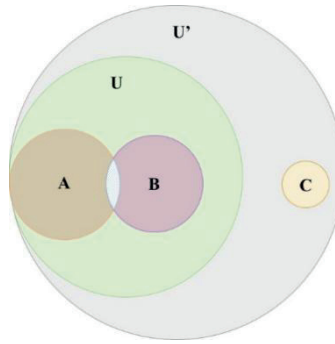


Figure 1. The sketch map of the technological complementarity definition

In order to quantitatively display the technological complementarity of a M&A, we need to design a measurement model. In this measurement model we use patents as a proxy for the technological scope of a company. As shown in Fig. 2 and Fig. 3, firm A is the acquiring firm (in green), firm B is the acquired firm (in purple), and U is the technological scope (in orange) which firm A plans to strategically develop. Under these circumstances, we propose to measure technological complementarity by evaluating the degree that firm B could complement the missing parts of firm A under the scope of U. We find this question to be equal to the one of measuring the technological similarity between firm B and the missing parts of firm A under U. Here, we denote the parts of U that are not in A as A' (the shadow area in yellow). We mark the common part between firm B and scope U but without the inner part of A as B' (the shadow area in purple). That is, $A'=U-(A\cap U)$, $B'=\{B-(A\cap B)\}\cap U$, and $B' \subset A'$. Therefore, the final question is to assess how B' could contribute to A'. The methodological contribution of this paper is to propose measures of similarity and complementarity that take into account technological distance.

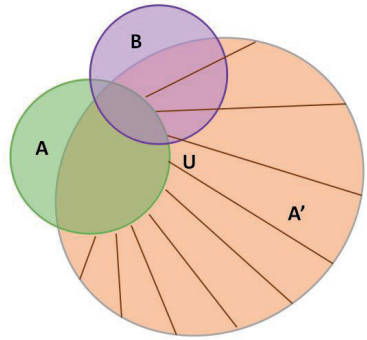


Figure 2. Diagram A': the potential complementarity space of A

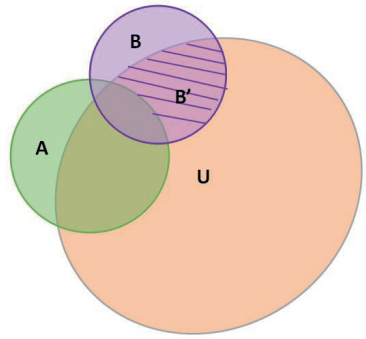


Figure 3. Diagram of B': the actual complementarity space of B over A

Data and Methodology

A co-occurrence matrix between IPC codes was produced for the patents in the A61 categories at the 7-digit level categorization using the Autumn 2017 version of the PATSTAT-database for the period 2000-1015. This resulted in a 188*188 matrix based on the co-occurrence of these 7-digit IPC codes between patent publications aggregated into

Inpadoc patent families. The International Patent Classification (IPC) was established by International Patent Classification Agreement in 1971, which provides a hierarchy of language-independent symbols for classifying patents and utility models according to different technical fields. To show the relationship between 7-digit IPC codes, we visualize the 7-digit IPC co-occurrence matrix using VOSViewer, following procedures similar to those used by Leydesdorff et al. (2014) at a 3 and 4-digit level, in Fig. 4.
The similarity between 7-digit IPC Codes i and j, is denoted as S_{ij}, and computed using the Ochiai Coefficient.

$$S_{ij} = \frac{C_{ij}}{\sqrt{C_i C_j}}$$

However, since the 7-digit IPC codes are not orthogonal (i.e. there is some similarity between them), we propose to use a similarity-weighted cosine instead of the conventional the cosine to calculate the similarity between firms. In doing so, we follow the same method used by Zhou et al. (2012) for comparing the disciplinary scope of countries. The proportion of patents in the 7-digit IPC code *i* of the acquiring firm A is denoted as A_i.The proportion of patents in the 7-digit IPC code *i* of the acquired firm B is given by B_i. Let $\phi(A,B)$ represent the normalized similarity between firms, which is defined as:

$$\phi(A,B) = \frac{\sum_{i,j=1}^{n} A_i B_j S_{ij}}{\sqrt{(\sum_{i,j=1}^{n} A_i A_j S_{ij})(\sum_{i,j=1}^{n} B_i B_j S_{ij})}}$$

Next, we calculate technological complementarity. As explained in the previous section, the complementary contribution of B to A, is only the space B'. Therefore, we can capture the complementarity with the vector B', which represents the number of patents of firm B in 7-digit IPC codes of the relevant scope U. In our example, we use only IPCs in the health area, i.e. in categories A61.

$$B' = \{B - (A \cap B)\} \cap U$$

When $B'_i = 0$, it indicates that B' does not have any knowledge under the specific 7-digit IPC code *i*. When $B'_i > 0$, it indicates that B' holds the technology on this specific knowledge base *i*, and the value of B'_i is the number of patents (a proxy of technology strength) of B'.
We use the similarity between 7-digit IPC codes S_{ij} to take into account the technological similarity between patent classes. The technological complementarity of an acquired firm B to an acquiring firm A can be estimated as:

$$TechCompl_{(B \to A)} = \sqrt{\sum_{i,j=1}^{n} B'^2_i S_{ij}}$$

Finally, the max-min method is used to normalise the technological complementarity.

Results: Case study

This paper selects technology-driven M&As in China healthcare industry as a case study following the introductions in 2011, a new policy for the healthcare industry introduced, the Medical Device GMP Certification. This policy raised the threshold for GMP certification, enhanced the technical barriers and spawned the tide of M&A in China healthcare industry. During the 12th five-year plan period (2011-15), the amount of M&A of China healthcare industry reached more than 150 billion Yuan (19.5 billion Euros).

For this paper, we collected the M&A events for the period 2013-2015 from CVSource (China Venture Source: http://www.cvsource.com.cn/). The cases were chosen according to the following criteria: the status of the M&A case was complete; the main party of the M&A was a company listed in Shenzhen and Shanghai Stock Exchanges; when the M&A event occurs, the main party had already been listed, ensuring that the company data could be checked. Then, 307 Chinese healthcare M&A cases come out to our list. Next, our study further selected Tech M&As. According to Ahuja and Katila (2001), we consider technology as a motivating factor for an acquisition or as a part of the transferred assets. Since in 2006 Chinese government required all M&As cases to mention the motivation in the M&As announcement documents, in our study, we carefully read the M&A announcement of each M&A event and selected the M&A case that explicitly refers to technical resources in the M&A Motivation. Here, the technical resources include patents, new products, research projects, R&D teams, etc. Finally, 24 cases were selected.

In our study, all categories of the A61 were assumed to make the U set. Since the main IPC of health care patents are distributed in A61 (Medical or Veterinary science and Hygiene), we use the 7-digit IPC codes of A61 as a proxy in our measurement model for the "technological knowledge base" of a company. Therefore, A61 defines the technological scope U of the firms for this study. This scope is shown in Fig. 4, which uses VOSviewer to visualized the A61 co-occurrence matrix.

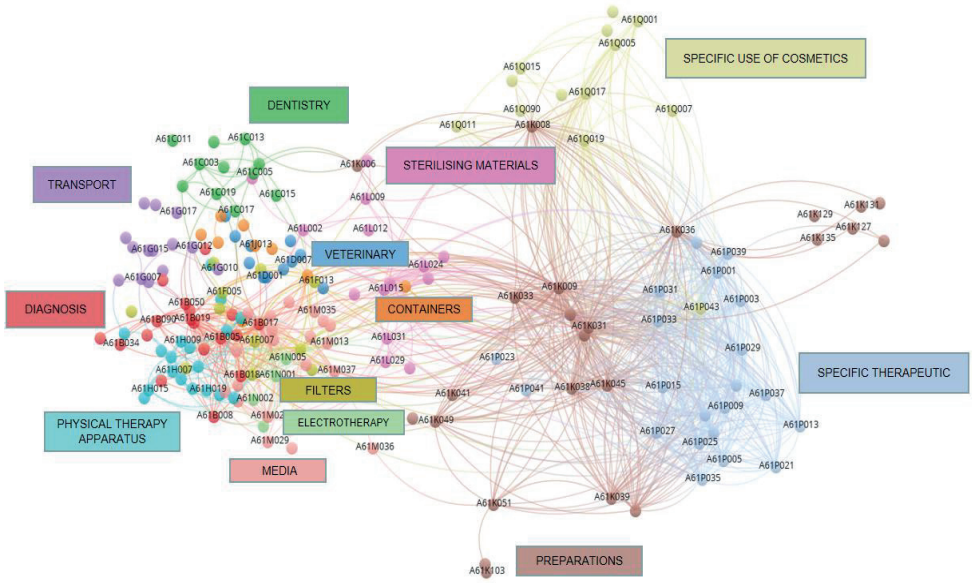


Figure 4. Technological similarity map of IPC codes for the space of A61 (health)

Figure 4 shows the labels of the most prominent clusters of classes in the study. The paper modifies the clustering classification. In Figure 4, we can see that there is a relatively close distance between DIAGNOSIS, DENTISTRY, VETERINARY, FILTERS, TRANSPORT, PHYSICAL THERAPY APPARATUS, CONTAINERS, MEDIA and ELECTROTHERAPY.

We show two cases of technology mergers as examples. The first example is case number 7 in Table 2. Table 2 shows the similarity and complementarity specific values of 24 M&A cases. On November 15, 2013, Tofflon acquired 46.15% stock in Divine. Tofflon's main research area is pharmaceutical machinery. Divine's main research area is biomedical materials. Tofflon thought that Divine's competitive advantage was sticky absorbable medical film. Divine was expected to speed up the registration of the series of products, and to form a product line based on sticky absorbable medical film after M&A. Tofflon's main IPC distribution was A61L 2/00 (Methods or apparatus for disinfecting or sterilising materials or objects other than foodstuffs or contact lenses; Accessories therefor). Divine's main IPC distribution is A61L 31/00 (Materials for other surgical articles). The IPC maps of Tofflon and Divine are shown in in Figure 5 and Figure 6. We can see that they work in complementary areas. Divine expanded Tofflon in the A61F area.

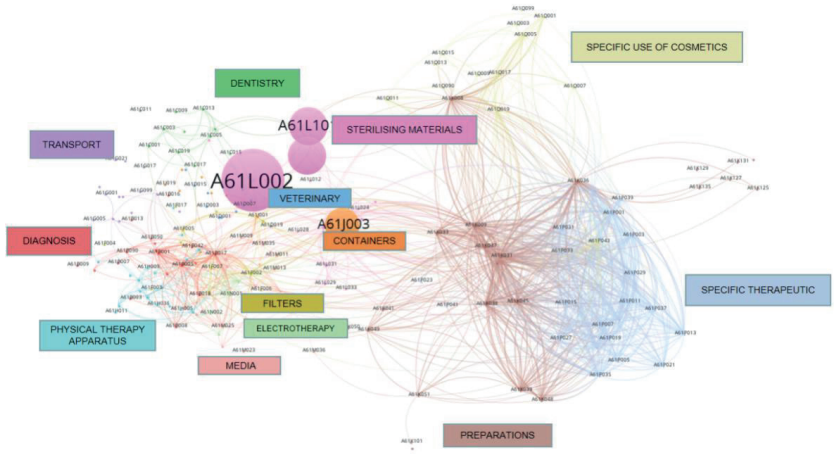


Figure 5. IPC overlay maps of Tofflon (acquiring firm)

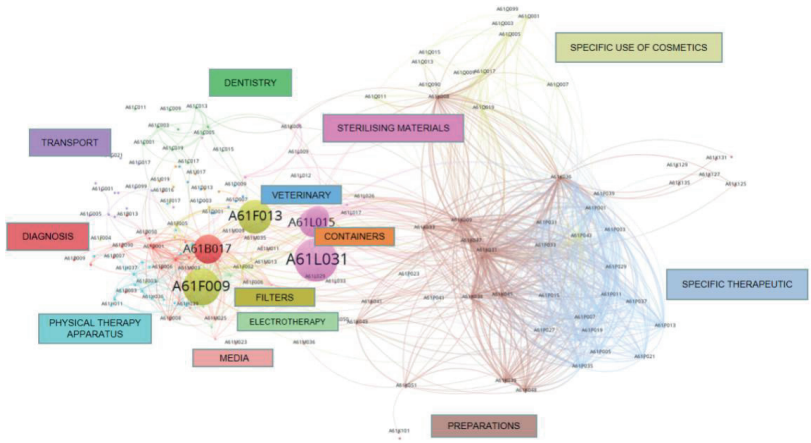


Figure 6. IPC overlay maps of Divine (acquired firm)

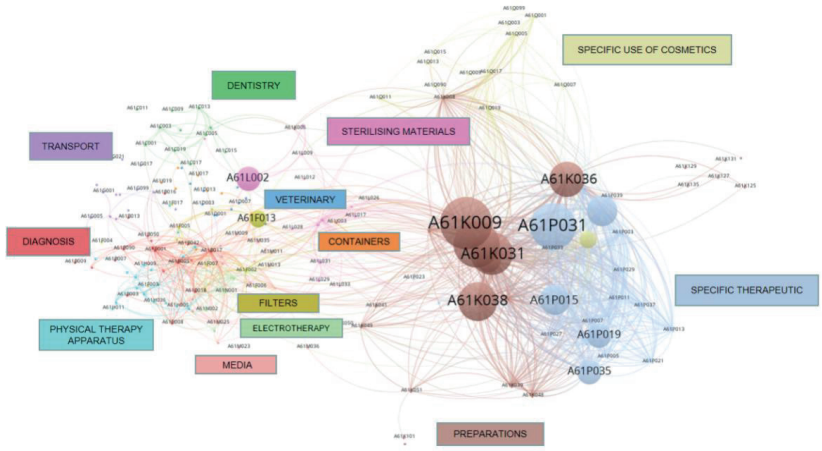


Figure 7. IPC overlay maps of Ankebio (acquiring firm)

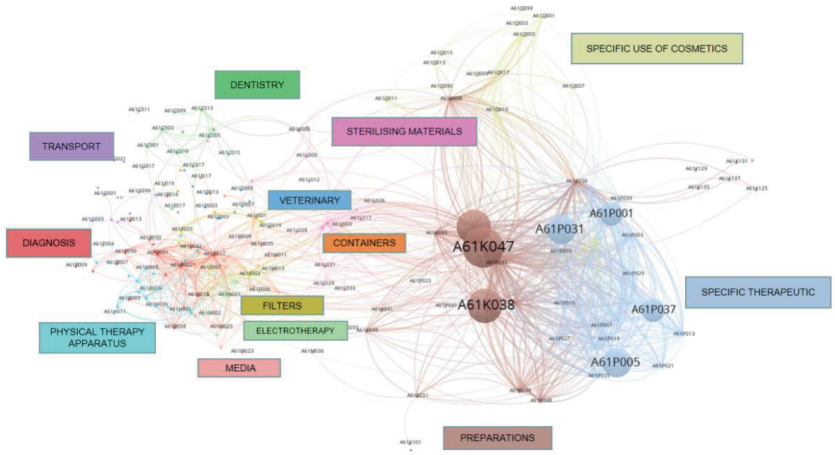


Figure 8. IPC overlay maps of Shanghai soho-yiming Pharmaceutical (acquired firm)

The second example is case number 18 in Table 2. On March 14, 2015, Ankebio acquired 100% stock of the company Shanghai soho-yiming Pharmaceutical. Ankebio's main research area was precision medicine. Shanghai soho-yiming Pharmaceutical's main research area was peptides. Shanghai soho-yiming Pharmaceuticals wanted Ankebio's competitive advantage in peptide API. The M&A event enabled Ankebio to expand its distribution of peptide drugs. In addition, Ankebio's main IPC was A61K 9/0 (Medicinal preparations characterized by special physical form). Shanghai soho-yiming Pharmaceutical's main IPCs are in A61K 47/00 (Medicinal preparations characterized by the non-active ingredients used, e.g. carriers or inert additives; Targeting or modifying agents chemically bound to the active ingredient). The IPC maps of Ankebio and Shanghai soho-yiming Pharmaceutical are shown in Figure 7 and Figure 8. We can see that they work in very similar areas, their main areas are concentrated in A61K and A61P – hence there is very low complementarity.

Finally, we calculated the similarity and complementarity of 24 M&A cases according to the methods discussed above, as shown in Table 2. A scatter plot of the two variables shown in Figure 9. The negative and moderate Pearson correlation observed (-0.63), confirms that although correlated, similarity and complementarity are two different phenomena.

Table 2. similarity and complementarity of 24 M&A cases.

Case	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*
similarity	0.03	0.94	1.00	0.43	0.90	0.83	**0.01**	0.00
complementarity	0.46	0.17	0.00	0.31	0.08	0.15	**0.28**	0.31

Case	*9*	*10*	*11*	*12*	*13*	*14*	*15*	*16*
similarity	0.99	0.99	0.89	0.00	0.00	0.53	0.50	0.00
complementarity	0.00	0.16	0.23	1.00	0.43	0.17	0.08	0.51

Case	*17*	*18*	*19*	*20*	*21*	*22*	*23*	*24*
similarity	0.31	**0.93**	0.58	0.98	0.87	0.77	0.04	0.62
complementarity	0.47	**0.07**	0.03	0.46	0.09	0.00	0.20	0.03

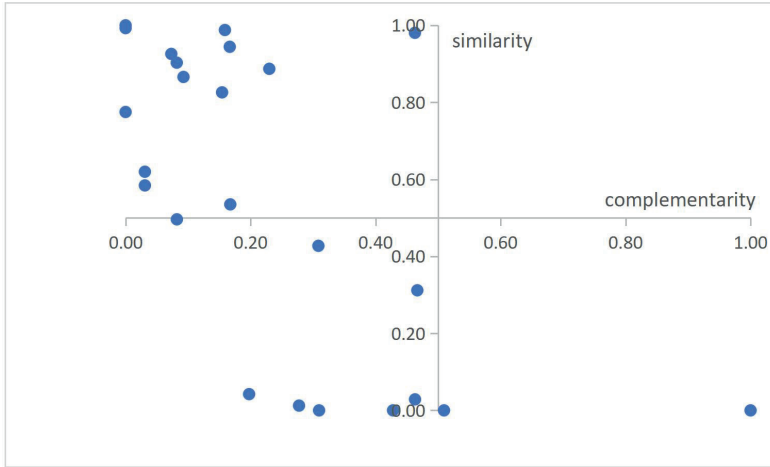


Figure 9. similarity and complementarity quadrant diagram for 24 M&A cases

Discussion

The paper focuses on a new conceptualization and calculation of complementarity. Following the proposed conceptualization of complementarity, the values of complementarity are not equal to one minus the similarity values. In other words, complementarity does not equate with dissimilarity. Previous definitions of complementarity were calculated using an orthogonal (i.e. rectangular) coordinate system. This paper uses an affine coordinate system for calculating the complementarity, which takes into account the intrinsic similarity between 7-digit IPC codes. Regarding the advantages of using the affine coordinate system, Zhou et al. (2012) and Zhou (2015) made a detailed introduction.

If companies want to expand into new areas, they need to acquire complementary companies. One can speculate that when the degree of complementarity and similarity are relatively high, mergers are positive. This means that in practice, complementarity may not go to maximum values, because the higher the complementarity, the higher the cost of integration after M&A. If there is a high similarity is between acquiring firm and acquired firm, the cost of integration may be low, but if this is also associated with low complementarity this means that the acquiring firm will not expand much their technological scope.

The purpose of the study is to help understand decision making processes for acquiring firms in M&A. If the acquiring firms have many choices in the merger and acquisition, mapping the degree of complementarity and similarity may help narrow down M&A targets. They can consider the technological spaces involved in the merger, the difficulty of post-merger integration and the monetary cost of the merger at the same time.

Acknowledgments

We acknowledge support from National Science Foundation of China (Grant No. 71774013).

References

Amesse, F., Latour, R., Rebolledo, C. & Séguin-Dulude, L. (2004). The telecommunications equipment industry in the 1990s: from alliances to mergers and acquisitions. *Technovation*24(11), 885-897.

Ahuja, G. & Katila, R. (2001). Technological acquisitions and the innovation performance of acquiring firms: a longitudinal study. Strategic Management Journal 22(3), 197−220.

Cantwell, J. & Vertova, G. (2004). Historical evolution of technological diversification. *Research Policy*33(3), 511-529.

Cassiman, B., Colombo, M.G., Garrone, P. & Veugelers, R. (2005). The impact of M&A on the R&D process: An empirical analysis of the role of technological- and market-relatedness. *Iese Research Papers* 34(2), 195-220.

Gutiérrez, Nicolás. (2016). The integration of complementary knowledge through collaboration among public R&D organisations: lessons from the agri-biotechnology innovation system in Uruguay. Doctoral thesis (PhD), University of Sussex. https://sro.sussex.ac.uk/65907/.

Heeley, Michael B., King, David R., Covin & Jeffrey G. (2006). Effects of firm R&D investment and environment on acquisition likelihood. *Journal of management studies*43(7), 1513-1535.

Hitt, M.A., Hoskisson, R.E., Ireland, R.D. & Harrison, J.S. (1991). Effects of Acquisitions on R&D Inputs and Outputs. *Academy of management journal*34(3), 693-706.

Hussinger, K. (2010). On the importance of technological relatedness: SMEs versus large acquisition targets. *Technovation* 30(1), 57-64.

James, A.D., Georghiou, L. & Metcalfe, J.S. (1998). Integrating technology into merger and acquisition decision making. *Technovation* 18(8-9), 563-573.

Leydesdorff, L., Kushnir, D. & Rafols, I. (2014). Interactive overlay maps for US patent (USPTO) data based on International Patent Classification (IPC). *Scientometrics* 98(3), 1583-1599.

Makri, M. & Lane, P.J. (2010). Complementary technologies, knowledge relatedness, and invention outcomes in high technology mergers and acquisitions. *Strategic Management Journal*31(6), 602-628.

Paruchuri, S. & Eisenman, M. (2012). Microfoundations of Firm R&D Capabilities: A Study of Inventor Networks in a Merger. *Journal of management studies*49(8), 1509-1535.

Ranft, A.L. & Lord, M.D. (2002). Acquiring New Technologies and Capabilities: A Grounded Model of Acquisition Implementation.*Informs*13 (4), 420-441.

Tingting Ma, Yi Zhang, Lu Huang, Lining Shang, Kangrui Wang, Huizhu Yu & Donghua Zhu. (2017). Text mining to gain technical intelligence for acquired target selection: a case study for china's computer numerical control machine tools industry. *Technological Forecasting & Social Change*, 116, 162-180.

Williamson, O.E. (1983). *Markets and hierarchies.* New York: Free Press.

Wulf, J. & Singh, H. (2013). How Do Acquirers Retain Successful Target CEOs? The Role of Governance. *Management Science*57(12), 2101-2114.

Xiao Zhou, Liliana Mitkova, Yi Zhang, Lu Huang, Scott Cunningham, Lining Shang, Huizhu Yu & Kangrui Wang. (2018). Technology-driven mergers and acquisitions of Chinese acquirers: development of a multi-dimensional framework for post-innovation performance. *International Journal of Technology Management*78(4), 280-309.

Zhou, Q., Rousseau, R., Yang, L., Yue, T., & Yang, G. (2012). A general framework for describing diversity within systems and similarity between systems with applications in informetrics. *Scientometrics93*(3), 787-812.

Zhou Q. (2015). Theoretical research and application experiment on the affine vector space model for co-occurrence analysis. Doctoral thesis (PhD), University of Chinese Academy of Sciences. http://dpaper.las.ac.cn/Dpaper/detail/detailNew?paperID=20010695.

Author-selected Keyword Semantic Function Analysis: A Case Study of Informetrics

Zhifeng Liu[1], Xin Li[2], Qikai Cheng[3] and Wei Lu[4]

[1]*zhifengliu@whu.edu.cn*
School of Information Management, Wuhan University, Wuhan(China)
Information Retrieval and Knowledge Mining Laboratory, Wuhan University, Wuhan(China)

[2]*lucian@whu.edu.cn*
School of Information Management, Wuhan University, Wuhan(China)
Information Retrieval and Knowledge Mining Laboratory, Wuhan University, Wuhan(China)

[3]*chengqikai@whu.edu.cn*
School of Information Management, Wuhan University, Wuhan(China)
Information Retrieval and Knowledge Mining Laboratory, Wuhan University, Wuhan(China)

[4]*weilu@whu.edu.cn*
School of Information Management, Wuhan University, Wuhan(China)
Information Retrieval and Knowledge Mining Laboratory, Wuhan University, Wuhan(China)

Abstract

Keywords for scientific literature are terms selected and created by authors, and are, in general, considered a core element that summarizes and represents the papers' content, which are often used for the analysis of research hotspots and trends. Keyword semantic function means the semantic role or specific function that a keyword plays in a scientific literature. The purpose of this study is to build a classification framework and a dataset for keyword semantic function analysis. First, we develop a classification framework of author-selected keyword semantic function in the field of informetrics, including *research topic, research method, research object, research area, data,* and *others.* Then, the data obtained from the *Journal of Informetrics* (JOI) and *Scientometrics* were manually annotated. Thereafter, the descriptive analysis of annotation dataset is carried out. The result shows that the dataset includes 2823 papers and 13975 author-selected keywords, and it was revealed that the largest proportion of semantic function in the dataset is *research method*, followed by *research topic*. Finally, we analyzed the high-frequency author-selected keywords with semantic functions of *research topic*, *research method* and *data* to see the landscape in the field of informetrics. Furthermore, the dataset in this study can be applied in academic retrieval and scholarly recommendation, etc.

Introduction

Recently, the scientific knowledge has increased rapidly with the rapid progress of scientific research (Bergstrom, Börner, & Fortunato, 2018). It is more difficult for researchers to obtain the required papers in a large number of academic papers and to find research hotspots and trends to meet the information needs in the research process. The research on author-selected

keywords can help to solve the above problem due to the fact that author-selected keywords of academic papers usually represent the content of academic papers (Kwon, 2018). The current studies on author-selected keywords are mainly focused on keyword frequency analysis, keyword co-occurrence analysis, information retrieval, citation recommendation (Olmeda-Gómez, Ovalle-Perandones, & Perianes-Rodríguez,2017; Xu & Liu,2018), etc. However, these studies that only based on the simple statistical analysis and matching of keywords failed to consider from the semantic level of keywords, leading to deviations in analysis or matching results, so it is difficult to meet the information needs of researchers.

In fact, authors select keywords for their purpose. The author-selected keywords usually indicate the areas and objects of their research, and reveal the research topics, and describe the methods used in the research, that is, author-selected keywords perform a certain semantic function in academic papers. The research on semantic functions of author-selected keywords can be applied to many fields, such as informetrics, information retrieval and citation recommendation. There have been a few studies highlighting the semantic function of author-selected keywords. Hu and Chen (2014) divided the semantic function types of author-selected keywords into research topics, fields, scopes, theoretical methods and sub-knowledge points. In his work, author-selected keywords were manually annotated to explore the influence of keyword semantic features on co-word analysis. Liu et al. (2016) divided the author-selected keywords of academic papers on big data research in the field of Library and Information Science into research topic, research method, tool and technology. By annotating a small number of author-selected keywords manually, he studied the dynamic evolution process of big data in the field of Library and Information Science multi-dimensionally. From the above discussion, however, there are still some shortcomings in the research of keyword semantic function. First of all, the academic community has not yet reached an agreement on the semantic function classification of author-selected keywords from scientific literature. Then, there is no standard dataset to support the research of keyword semantic function. Last but not least, the keyword semantic functions have not been applied widely in the field of informetrics.

This study attempts to develop a classification framework for the semantic function of scientific literature in the field of informetrics through literature research and dataset research. Author-selected keywords in *Journal of Informetrics* and *Scientometrics* are annotated, and the annotation dataset is constructed to support subsequent research on semantic analysis and understanding of academic papers. Then, the semantic function distribution characteristics of the dataset are revealed. Finally, we analyze the content of the high-frequency author-selected keywords with different semantic functions.

The definition of semantic function of author-selected keywords

Academic papers are not only one of the main forms of scholars' research results but the main media of academic exchanges, a certain research in which mainly includes research background, research objects, research questions, theoretical basis, research methods, tools and research conclusions (Cheng, 2015). In order to facilitate researchers' understanding of the content of academic papers and meet the retrieval needs, most journals require authors to provide keywords for their academic papers. Author-selected keywords have rich semantic information, meaning that they own different semantic functions. They are the refinement of an academic

paper, which can reflect the paper's content very well. In this study, the semantic function of author-selected keywords refers to the cognitive and understanding of author-selected keywords from a semantic perspective, which is the corresponding content or function in the academic text environment.

The following example illustrates the semantic functions of author-selected keywords. As shown in Figure 1, there are three author-selected keywords in this paper: "Scientometrics", "Public Research Institutions" and "Scientific Performance". According to the title and abstract of this paper, it can be inferred that this paper puts forward a new indicator to evaluate the performance of public research institutions. Thus, we can deduce the semantic functions of the three given author-selected keywords. The research of this paper belongs to the field of Scientometrics, so the semantic function of the keyword "Scientometrics" in this paper is the *research area*; the object of this study is the public research institutes, so the semantic function of the keyword "Public Research Institutions" is the *research object*; the problem to be solved is the evaluation of the performance of public scientific research institutions, so the semantic function of the keyword "Scientific Performance" is the *research topic*. Note that the same keyword may have different semantic functions in different academic papers and vice versa.

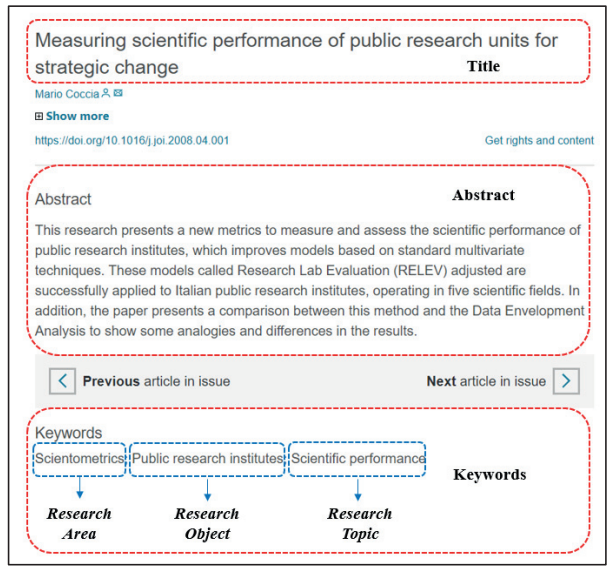


Figure 1. An example of keyword semantic functions in an article.

Methodology

We have followed a set of procedures to analysis semantic functions of author-selected keywords in the field of informetrics. For this, the methodology shown in Figure 2 has been established and is described below step by step.

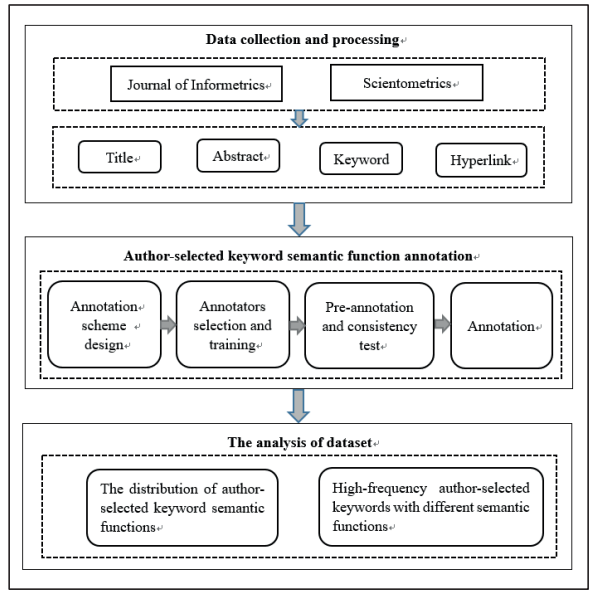


Figure 2. Methodology of analysis.

Data collection and processing

In this study, *Journal of Informetrics* and *Scientometrics* were selected as data sources for annotation. The link, title, abstract and author-selected keywords of every paper published in 2007-2017 were obtained from the official websites of *Journal of Informetrics* and *Scientometrics*. However, papers published in *Scientometrics* included is only from 2010 to 2017 because *Scientometrics* provide author-selected keywords after 2009. The final annotation data of 2823 papers are shown in Table 1.

Table 1. Overview of annotation data.

Name	*Journal of Informetrics*	*Scientometrics*	*Total*
Total number of papers	693	2130	2823
Total number of author-selected keywords	3310	10665	13975
Average number of keywords per article	4.7763	5.007	4.9504

The time distribution of papers and author-selected keywords on *Journal of Informetrics* and *Scientometrics* are presented in Figure 3. The trend of the number of author-selected keywords in the two journals is the same as the trend of the number of papers published in the two journals. The trend of the *Journal of Informetrics* is increasing at the beginning, then decreasing, and finally stabilizing. On the contrary, the beginning stage of *Scientometrics* has a decreasing trend with an increasing trend in the later stage. The time distribution of the average number of author-selected keywords per article on the *Journal of Informetrics* and *Scientometrics* are presented in Figure 4. There are some fluctuations in the average number of author-selected keywords per article of the two journals, but the overall trend is increasing over time.

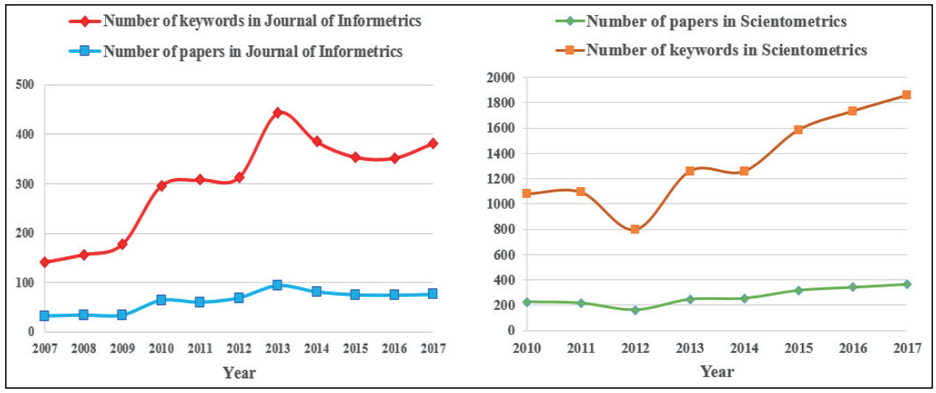


Figure 3. Time distribution of papers and author-selected keywords on the *Journal of Informetrics* and *Scientometrics*.

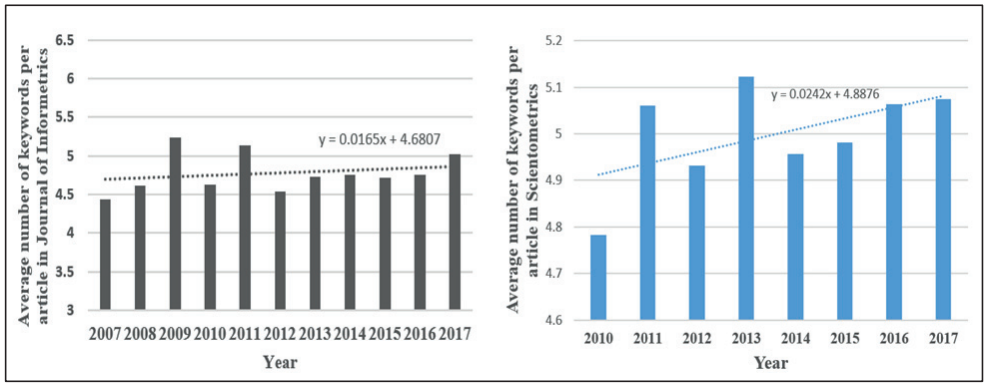


Figure 4. Time distribution of the average number of author-selected keywords per article on the *Journal of Informetrics* and *Scientometrics*.

Author-selected keyword semantic function annotation

Annotation scheme design

In the previous studies regarding term semantic function, words in scientific papers have been recognized as "Problem", "Solution", "Method", "Technology", "Application", "Dataset", "Other" etc. (Cheng, 2015; Heffernan & Teufel, 2018; Kondo, Nanba, & Takezawa,2009; Siddiqui, Xiang, & Parameswaran, 2016). Concerning the different semantic function of each author-selected keyword in each article, we present an annotation scheme for keywords based on empirical work in content analysis. In the first place, we captured all possible semantic functions of author-selected keywords. Then, to simplify our analysis, these semantic functions were integrated to a smaller set that comprised of five most frequent semantic functions. Specifically, the annotation scheme for semantic functions of author-selected keywords includes the following categories: (1) Research topic; (2) Research method; (3) Research object;

(4) Research area; (5) Data; and (6) Others. The detailed description and source for each category of semantic functions are shown in Table 2.

In order to guarantee the preciseness of keyword semantic function annotation, the method of human annotation is selected. The semantic function of author-selected keywords is difficult to annotate because, in principle, it requires interpretation of the author's intentions and the content of the entire paper. Consequently, in most cases, it is impossible to know its exact semantic function without understanding its academic context, since the same keyword can have a totally different semantic function in different conditions.

Annotators selection and training

Before semantic function annotating, two Ph.D. candidates were firstly selected from the School of Information management, Wuhan University. Four criteria were used in the selection of annotators. Specifically, the annotators had to: (1) be very familiar with the field of informetrics and bibliometrics; (2) have good English reading and writing skills; (3) have published more than two academic articles in peer-reviewed journals in the field of informetrics; and (4) be in or beyond their second year in the Ph.D. program. Then, the selected annotators were trained and asked to point to textual evidence they have had for assigning a particular semantic function.

Pre-annotation and consistency test

To guarantee annotation consistency, prior to starting the annotating, we randomly chose 69 articles (9.96%) comprising of 337 author-selected keywords from the JOI dataset and arranged for two annotators to parallelly annotate author-selected keyword functions. Then, the Jean Carletta's kappa coefficient, which is a statistic measuring pairwise agreements among a set of coders category judgments (Carletta, 1996), was used for quantifying the consistency. Finally, the coefficient was 0.830 (above 0.75), which was considered high enough for annotating to proceed separately, particular given the conservative nature of the kappa coefficient.

Annotation

In the processing of annotating, annotators were asked to carefully read the title and abstract for a comprehensive understanding of the academic context of each author-selected keyword in the original dataset and were encouraged to click the hyperlink for its full text to make a further confirmation. Moreover, annotators were asked to record the Annotation Confidence(ac) of each article. The value of $ac \in [1,2,3,4,5]$, in which a higher value of ac represents that the annotator is more confident in his or her work. If an article's value of ac is below 4, the article will be annotated again by both annotators together.

Table 2. The annotation scheme of semantic functions of author-selected keywords in the field of informetrics.

Types of semantic functions	*Descriptor*	*References*
Research Area	The academic area or research back -ground in the academic text.	Galán (2002)
Research Object	People, organization, institutions, etc.	Cheng (2015); Li (2017)
Research Topic	Topics or problems discussed in scientific literature.	Cheng (2015); Galán (2002); Hefferna et al. (2018); Kondo et al. (2009)
Research Method	Research method includes the indicators, theories, laws, models and algorithms used in the paper, etc.	Augenstein et al. (2017); Cheng (2015); Hefferna et al. (2018); Mesbah et al. (2017); Sahragard & Meihami (2016); Tsai et al. (2013)
Data	The dataset used in the paper or the source of the dataset.	Kondo et al. (2009); Mesbah et al. (2017); Sahragard & Meihami (2016)
Others	Can't be included in the former categories.	Kondo et al. (2009)

Results

The distribution of author-selected keyword semantic functions

Python was used to extract and count the semantic functions of author-selected keywords in the annotated dataset. As shown in Figure 5, the number of author-selected keywords with the semantic function of *research method* is the largest, followed by the author-selected keywords with semantic function of *research topic* accounting for 68.83%, probably because the author-selected keywords with the semantic function of *research method* and *research topic* can better represent the core content of academic papers. In addition, the least proportion of semantic function is *data* accounting for only 2.55%, which indicates that authors rarely present their research data source or dataset as the keywords for academic papers.

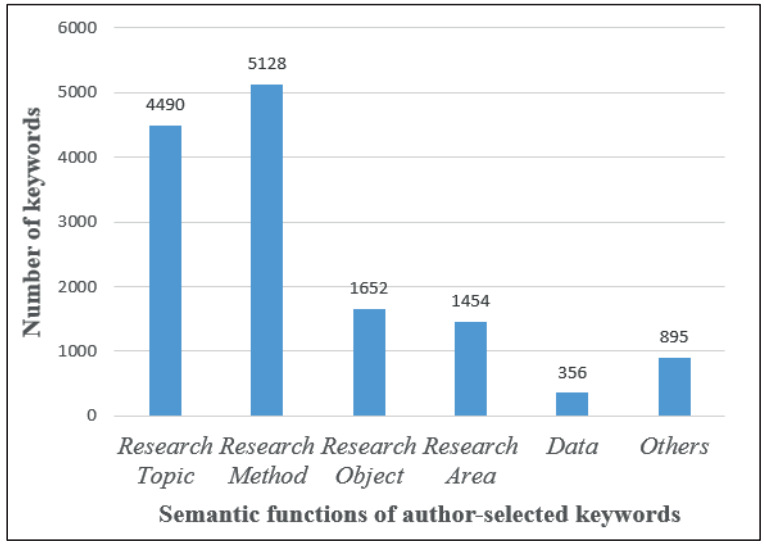


Figure 5. The distribution of author-selected keyword semantic functions.

High-frequency author-selected keywords with different semantic functions

The author-selected keywords can reflect the information that the author mostly wanted to express. The analysis of high-frequency author-selected keywords can show the recent research hotspots and trends. Thus, we analyzed the top 20 high-frequency author-selected keywords with semantic functions of *research topic*, *research method* and *data* respectively.

High-frequency author-selected keywords with the semantic function of research topic

Figure 6 shows the top 20 high-frequency author-selected keywords with the semantic function of *research topic*. These author-selected keywords throw light on the research topic in which scholars are engaged with. We analyzed the content of the author-selected keywords and classified them into six groups: **(1) Scientific evaluation and rank**, which includes Research Evaluation, Scientific Productivity, Peer Review, Rank, Research Performance, Research Productivity, Journal Rank, Productivity, University Rank, Research Output; **(2) Research collaboration**, which includes Scientific Collaboration, Collaboration, Research Collaboration, Co-Authorship, International Collaboration; **(3) Citation analysis,** which includes Citation, Citation Impact; **(4) Interdiscipline**; **(5) The detection of research hotspots and frontiers**; **(6) Knowledge flow and innovation diffusion**, etc.

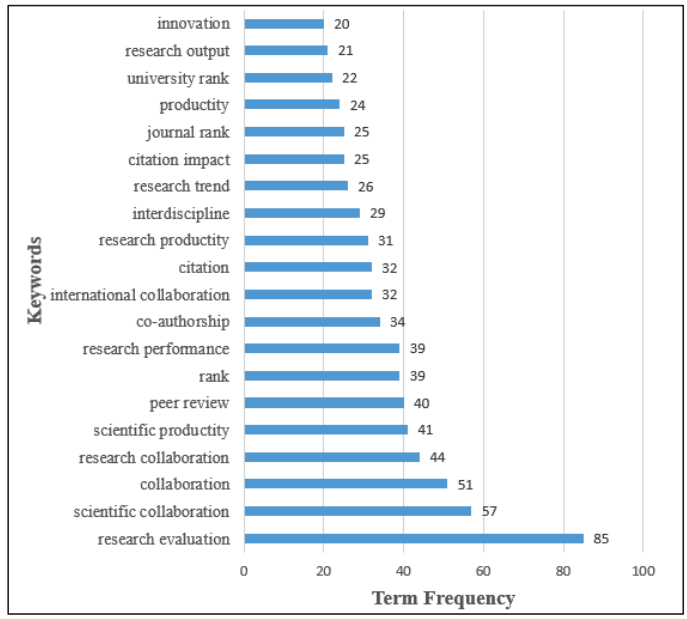


Figure 6. High-frequency author-selected keywords with the semantic function of *research topic*.

High-frequency author-selected keywords with the semantic function of research method

Research methods play an important role in a discipline. With the development of the field of informetrics, the number of its own research methods have been increasing. At the same time, it has been drawing on the research methods of other disciplines. As can be seen in the Figure 7, the main research methods in the field of informetrics are as follows: **(1) Indicators**, such as H-Index, Impact Factor, G-Index; **(2) Citation analysis**, such as Citation Analysis, Bibliograph Couple, Citation Network; **(3) Content analysis**; **(4) Text mining and visualization**, such as Text Mine, Co-Word Analysis, Cluster Analysis; **(5) Complex network analysis**, such as Social Network Analysis, Network Analysis, Citation Network, Collaboration Network; **(6) Patent analysis**.

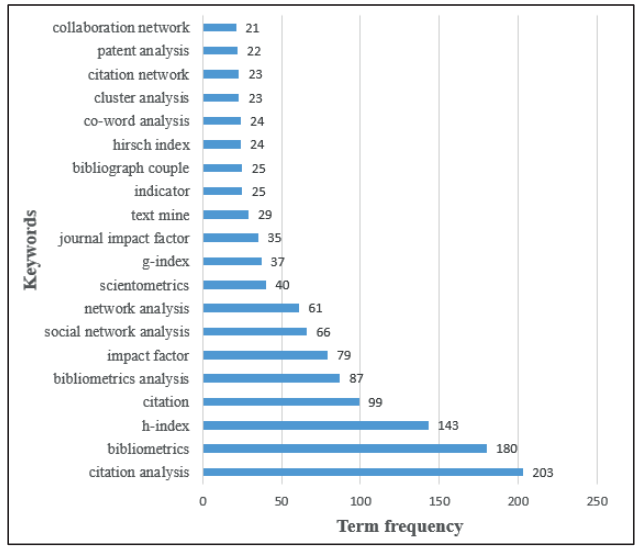


Figure 7. High-frequency author-selected keywords with the semantic function of *research method*.

High-frequency author-selected keywords with the semantic function of data

With the continuous development of technology, the amount of generated data and the types of data are increasing. How to make good use of these multi-source heterogeneous data has brought opportunities and challenges to the research in the field of informetrics. Statistical analysis of author-selected keywords with semantic function of *data* can help us to understand the data sources and types of data used in the field of informetrics.

Figure 8 shows that citation databases and academic search engines are the main data sources in the field of informetrics. In addition, with the rapid development of altmetrics, academic social network platforms and databases for altmetrics have become important data sources. The citation databases mainly include Web of Science, Scopus, Pubmed, etc. Search engines mainly include Google Scholar, Academic Search Engines, etc. Academic social network platforms and databases for altmetrics mainly include Mendeley, F1000, etc. The data obtained from these data sources mainly include metadata, reference data, usage data (such as the download data, browsing data, etc.), etc. It implies the fact that the data sources and the types of data used in the field of informetrics are constantly enriched with the rise of the mobile internet and social media.

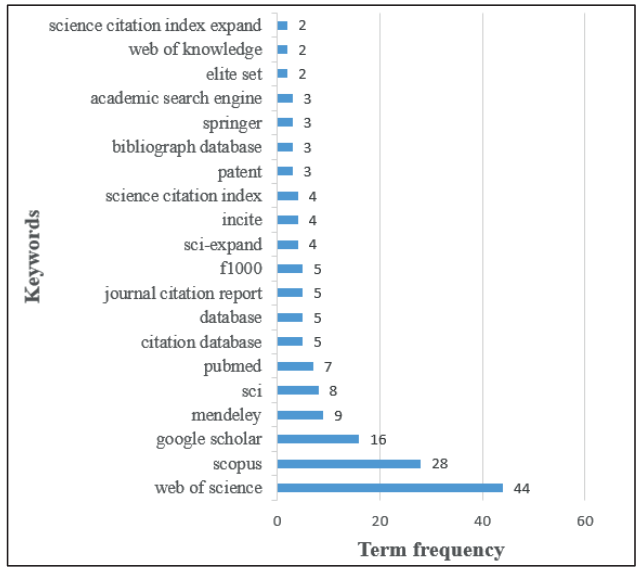


Figure 8. High-frequency author-selected keywords with the semantic function of *data*.

Conclusion and future work

From the perspective of semantic functions of author-selected keywords from scientific literature, this study constructed a classification framework for semantic functions of author-selected keywords in the field of informetrics. Based on this framework, a dataset of semantic function annotated author-selected keywords was constructed, which is of certain theoretical and practical value. Eventually, by analyzing the annotated dataset, the distribution characteristics of the semantic functions of author-selected keywords were revealed, which at the same time, can systematically present the research status of informetrics, which can benefit not only researchers in terms of promoting understanding of the entire field, but also governments for funding agencies, as well as provide some insights into the application of this dataset.However, there are several shortcomings in this study. Because the process of annotating of the semantic functions of author-selected keywords is difficult and time-consuming, only the author-selected keywords on *Journal of Informetrics* and *Scientometrics* are selected as samples, which are relatively limited. In the future, we will explore the automatic annotation of author-selected keywords' semantic functions so that the size of the sample could be expanded and a larger semantic function annotation dataset can be built. In addition, we will further explore the application of the dataset in informetrics, academic retrieval and citation recommendation.

Acknowledgments

This work was funded by the Major Project of the National Social Science Foundation of China (17&ZDA292) and the National Natural Science Foundation of China (71473183).

References

Augenstein, I., Das, M., Riedel, S., & Vikraman, L. (2017). SemEval 2017 Task 10: ScienceIE - Extracting Keyphrases and Relations from Scientific Publications. Proceedings of the 11th International Workshop on Semantic Evaluations. Vancouver ： Association for Computational Linguistics ,546-555.

Carletta, J. (1996). Assessing Agreement on Classification Tasks: The Kappa Statistic. Computational Linguistics, 22(2), págs. 249-254.

Cheng, Q. (2015). Term Function Recognition of Academic Text. Wuhan: Wuhan University.

Fortunato, S., Bergstrom, C. T., & Börner, K. (2018). Science of science. Science, 359(6), eaao0185.

Galán, A. D. (2002). Textual Interaction: An Introduction to Written Discourse Analysis, by Michael Hoey. Atlantis, 24(2), 259-264.

Heffernan, K., & Teufel, S. (2018). Identifying problems and solutions in scientific text. Scientometrics, (1),1-16.

Hu, C., & Chen, G. (2014). Keyword characteristics of scientific papers and their impact on co-word Analysis. Journal of The China Society for Scientific and Technical Information, 33 (1), 23-32.

Kondo, T., Nanba, H., & Takezawa, T. (2009). Technical trend analysis by analyzing research papers' titles. Human language technology: challenges for computer science and linguistics．Berlin: Springer, 512-521.

Kwon, S. (2018). Characteristics of interdisciplinary research in author keywords appearing in Korean journals. Malaysian Journal of Library & Information Science, 23(2), 77-93.

Li X., Cheng Q. K., & Lu W. (2017). CS-LAS: A Scientific Literature Retrieval and Analysis System Based on Term Function Recognition (TFR). Proceedings of the 16th International Conference on Scientometrics and Informetrics (pp.1346-1356). Wuhan, China.

Liu, Z., Wang, X., & Bai, R. (2016). Research on Visual Analysis Method of Subject Evolution from Multi-Dimensional Perspective --- Taking the Big Data Research in the Field of Library and Information in China as an Example. Library and Information Service, 42 (06),67-84.

Mesbah, S., Fragkeskos, K., & Lofi, C. (2017). Facet Embeddings for Explorative Analytics in Digital Libraries. International Conference on Theory and Practice of Digital Libraries. Greece: Springer ,86-99.

Olmeda-Gómez, C., Ovalle-Perandones, M. A., & Perianes-Rodríguez, A. (2017). Co-word analysis and thematic landscapes in Spanish information science literature, 1985–2014. Scientometrics, 113(1), 195-217.

Sahragard, R., & Meihami, H. (2016). A diachronic study on the information provided by the research titles of applied linguistics journals. Scientometrics, 108(3),1-17.

Siddiqui, T., Xiang, R., & Parameswaran, A. (2016). FacetGist: Collective Extraction of Document Facets in Large Technical Corpora. ACM International on Conference on Information and Knowledge Management. ACM, 871-880.

Tsai, C. T., Kundu, G., & Roth, D. (2013). Concept-based analysis of scientific literature. Proceedings of the 22nd ACM international conference on Information & Knowledge Management (CIKM '13). ACM, New York, NY, USA, 1733-1738.

Xu, J., & Liu, Y. (2018). A bibliometric analysis for global research trends on ectomycorrhizae over the past thirty years. The Electronic Library, 36(4), 733-749.

Toward Better Growth Policies in a Modern Economy: The Comparison of Three Complexity Indices

Inga Ivanova,[1] Nataliya Smorodinskaya[2] & Loet Leydesdorff[3]

[1] inga.iva@mail.ru
National Research University Higher School of Economics (NRU HSE), Institute for Statistical Studies and Economics of Knowledge Myasnitskaya St. 20, 101000, Moscow (Russia)

[2] smorodinskaya@gmail.com
Institute of Economics of the Russian Academy of Sciences, Dept for Innovation Economy and Industrial Policy, Nakhimovsky Pr 32, 117218 Moscow (Russia)

[3] loet@leydesdorff.net
University of Amsterdam, Amsterdam School of Communication Research (ASCoR) PO Box 15793, 1001 NG Amsterdam (Netherlands)

Abstract

In modern economies transiting to innovation-driven systems, the evaluation of competitive advantages and growth potentials is increasingly measured using complexity. Two recent complexity measures—the Economic Complexity index (ECI) and the Fitness and Complexity index (FCI)—assume bipartite country-product network data. We provide the Modified Ecosystem Complexity Index (MECI) as a possible refinement to be considered in national economic growth policies. MECI is based on Lotka-Volterra equations, implying an ecosystem's approach to measuring complexity. We compare the three complexity measures using empirical data for 41 countries and shifting from the Revealed Comparative Advantage index to the Revealed Effectiveness Advantage index. The regression analysis shows that with respect to GDP per capita growth, the predictive power of the three measures improves when changing from RCA to REA index, while at the same time the difference among correlations of the three measures with initial diversity score and GDP per capita is reduced. MECI exceeds ECI and FCI with respect to growth prediction and correlation with GDP per capita Furthermore, MECI can be elaborated further by incorporating the technological complexity data as a third dimension.

Introduction

The modern post-industrial economy, usually referred to in the economic literature as an innovation-driven economy, is a fundamentally more complex system than the industrial one. The increase in complexity is due to technological progress and, in particular, the ICT revolution, have significantly modified the key development parameters of economic systems - their organizational structure, the composition of production factors, the mechanism of sustainable growth (Russell & Smorodinskaya, 2018). The industrial economy can be described as a hierarchic system with a centralized governance and linear model of development, which relies on massive investment activity and accumulation of capital. The post-industrial economy displays non-linear dynamics, relying on continuous innovation activity and accumulation of knowledge. By its properties, the latter belongs to the class of complex systems, whose behavior is described in the interdisciplinary theory of complex adaptive systems (the CAS theory) (Axelrod & Cohen 2000; Holland, 2002).

Since ecosystems demonstrate a complex, emergent, and non-predictable behavior, modern economies transiting to innovation-driven growth are no longer subordinate to traditional methods of forecasting and planning (OECD, 2015). A new, modern way of assessing the economies development prospects is concerned with measuring their complexity. To assess the potential of modern economies for sustainable growth under the dynamic upgrading of technologies and a sharply raised uncertainty, several indicators for measuring their functional complexity were proposed, in particular, the Economic Complexity Index (ECI) (Hidalgo & Hausmann, 2009) and the Fitness and Complexity index (Tacchella *et al.*, 2013). The first

indicator relies solely on linear relations between the estimated economic parameters. The second indicator entertains non-linear relations.
Against this background, we propose a new ecosystem based indicator - the Modified Ecosystem Complexity Index. In this technique, we model the pattern of correlations between all estimated economic parameters on similar nonlinear basis which is typical for network interactions among co-evolving components of complex systems (economic agents and their groups), thus reproducing the way of their self-organization. If modern economies are increasingly organized as non-linear network ecosystems, then indicators for assessing their complexity should be similarly constructed. In other words, we try to embed the ecosystem's approach into the process of measuring the economic complexity.
As a conceptual basis for justifying the ecosystem's approach to measuring complexity, we rely on the CAS theory, in regard to its description of organizational structure and process of self-organization of complex systems. As a methodological basis, we rely on the Method of Reflections and build its non-linear version. We test and compare all the three indicators in order to identify possible advantages of the modified index.
The paper further falls in three sections. Section 2 reviews merits and limitations of the two existing complexity indicators and introduces a detailed methodology of building the Modified Ecosystem Complexity Index. Section 3 describes the operationalization of the proposed complexity indicator, present data sources, obtained results of calculations, and their discussion. Section 4 contains concluding remarks and an outlook for future research. Additional information, comprising comparative countries' ranking under three different measures of economic complexity, is provided in the Appendix.

Method

One of the practical applications of the CAS theory is concerned with the estimation of development prospects of concrete economic systems through measuring their complexity.
As stems from Porter's theory of competitiveness (Porter, 1990) and is usually taken in modern economic literature, the development prospects of a territory correspond to its competitive advantages. According to Porter's theory, competitive advantages of a territory are associated not so much with the variety of products it can produce and export but rather with the development potential it can ever realize. In its turn, the development potential largely depends on the existing opportunities for economic growth, availability of advanced technologies, the achieved level of R&D, the quality of human capital, etc. In this sense, competitive advantages of a territory can be seen as its potentially existing but not yet realized additional development options. As applied to a network-based ecosystem, the number of possible pathways, or additional options of its development is determined by intensity and synergy of interactions between its actors and components (Leydesdorff, Johnson, & Ivanova, 2018). This implies that the more complex an economy is, the greater number of additional options (which can potentially be realized) it generates, and therefore, the better are its development prospects. In other words, one can estimate the development prospects of a given economy by means of measuring its complexity as an ecosystem.
The relevant methodology has been firstly introduced in the end of 2000s by American economists C. Hidalgo and R. Hausmann in the form of Economic Complexity Index (ECI). These authors proposed to determine economic complexity of individual countries according to complexity and diversification of their exported goods (Hidalgo & Hausmann, 2009). The focusing on national exports as a basic parameter considers the factor of globalization (almost all innovative goods and services are now produced in traded industries) and the modern complex model of production, when the process of creating new final goods has ultimately gone beyond national borders and got distributed, both geographically and functionally,

across the nodes of global value chains (Baldwin, 2013). Correlation of a country's economic complexity with complexity of its exports has got empirical confirmations (Hausmann, Hidalgo *et al*., 2013).

The point of departure in constructing ECI was the self-organization of CAS, viewed as the process comprised of multiple steps forming an iterative sequence. Building on this sequence, Hidalgo & Hausmann proposed to measure relative economic complexity of countries through a linear iterative procedure, known as Method of Reflections (MR). Using the method one can obtain ECI which correlates with a country's GDP per capita and therefore, can be used to predict its future growth (Kemp-Benedict). Meanwhile the value of variables used to measure economic competitiveness by means of ECI deviate from the initial diversity value in the course of iterations, which implies that successive iteration terms can hardly be interpreted.

In early 2010s, a group of Italian economists, Tacchella and colleagues (Tacchella *et al.,* 2013), while following Harvard's experience of constructing ECI, proposed another technique for measuring complexity, known as the Fitness and Complexity index (FCI). This indictor also deals with complexity of national export portfolio but constitutes a non-linear modification of the Method of Reflections. Relying on empirical observations, the authors assume that countries can export national products in accordance with their objective available capabilities. So that developed countries are able to manufacture and export all range of products, while less developed ones, only limited product range.

FCI can as well be used for assessing the potential of economic growth along with ECI. The advantage of FCI over ECI is that a high correlation of a country's economic complexity with initial diversity of its exports can be maintained at each stage of iteration.

Considering merits and limitations of both ECI and FCI, we propose the Modified Ecosystem Complexity Index (MECI), constituting a non-linear modification of the Method of Reflections. Here we integrate the ecosystem's approach into measuring complexity, keeping in mind that if economies are organized as non-linear network ecosystems, then indicators for assessing their complexity should have a similar construction.

Our method is based on the Method of Reflections used by Hidalgo & Hausmann (2009), and defined by the following iterative sequence:

$$\begin{aligned} k_{p,n} &= \frac{1}{k_{p,0}} \sum_{c=1}^{N_c} M_{c,p} k_{c,n-1} \\ k_{c,n} &= \frac{1}{k_{c,0}} \sum_{p=1}^{N_p} M_{c,p} k_{p,n-1} \end{aligned} \tag{1}$$

where:

$$\begin{aligned} k_{p,0} &= \sum_{c=1}^{N_c} M_{c,p} \\ k_{c,0} &= \sum_{p=1}^{N_p} M_{c,p} \end{aligned} \tag{2}$$

Matrix $M_{c,p}$ is obtained with respect to Balassa's (1965) Revealed Comparative Advantage (RCA) index:

$$M_{c,p} = \begin{cases} 1 \; if \; RCA_{c,p} \geq 1 \\ 0 \; if \; RCA_{c,p} < 1 \end{cases} \tag{3}$$

$$RCA_{c,p} = \frac{X_{c,p} / \sum_p X_{c,p}}{\sum_c X_{c,p} / \sum_{c,p} X_{c,p}} \tag{4}$$

Where $X_{c,p}$ stands for a Country (c)–Product (p) Exports matrix. ECI is defined as a limit of iterations of the vector $\vec{k}$:

$$\vec{k} = \lim_{n\to\infty} k_{c,n} \tag{5}$$

according the formula:

$$ECI = \frac{\vec{k} - <\vec{k}>}{stdev(\vec{k})} \tag{6}$$

The Method of Reflections can be as well extended to the technological domain (Ivanova, Strand, & Leydesdorff, 2018), incorporating the tripartite network of countries, technologies, and products. Non-linear generalization of the MR can be defined by the following set of equations:

$$\begin{aligned} k_{c,n} &= \frac{1}{k_{c,0}} \sum_{p=1}^{N_p} \sum_{t=1}^{N_t} \mathfrak{M}_{c,p,t} k_{p,n-1} k_{t,n-1} \\ k_{p,n} &= \frac{1}{k_{p,0}} \sum_{c=1}^{N_c} \sum_{t=1}^{N_t} \mathfrak{M}_{c,p,t} k_{c,n-1} k_{t,n-1} \\ k_{t,n} &= \frac{1}{k_{t,0}} \sum_{c=1}^{N_c} \sum_{p=1}^{N_p} \mathfrak{M}_{c,p,t} k_{p,n-1} k_{c,n-1} \end{aligned} \tag{7}$$

Eqs. (7) can be considered as a discrete version of generalized *Lotka-Volterra equations*, widely used in biological studies to describe the evolution of ecosystems. Though, as noted above, self-organization of economic ecosystems is driven by the generation and adaptation of new knowledge, which makes a clear difference with the mechanism of biological evolution (Beinhocker, 2006), we assume that application of Lotka-Volterra equations can be relevant for our purposes.

System (7) can be reduced to two dimensions. We do this in order to obtain a more objective comparison among ecosystem based approach and the other two complexity measures. Excluding the technology dimension from system (7) one can get:

$$\begin{aligned} k_{p,n} &= \frac{1}{k_{p,0}} \sum_{c=1}^{N_c} M_{c,p} k_{c,n-1} k_{c,n-1} \\ k_{c,n} &= \frac{1}{k_{c,0}} \sum_{p=1}^{N_p} M_{c,p} k_{p,n-1}\, k_{p,n-1} \end{aligned} \tag{8}$$

In essence, Eqs. (8) represent a non-linear generalization of Eqs. (1). So, we define MECI, as calculated according to Eqs. (8), in the same way as we defined ECI in Eq. (6). Following Tacchella *et al.* (2013), we also calculate the Fitness ($F_c^{(n)}$) index, as defined by the following iterative sequences:

$$\tilde{F}_c^{(n)} = \sum_p M_{cp} Q_p^{(n-1)}$$

$$\tilde{Q}_p^{(n)} = \frac{1}{\sum_c M_{cp}\left(1/F_c^{(n-1)}\right)} \tag{9}$$

where $Q_p^{(n)}$ is the Product Complexity Index which measures complexity of separate product items by comparing countries exporting the same products

At each step of the iteration, the intermediate values are first computed and then normalized as follows:

$$F_c^{(n)} = \frac{\tilde{F}_c^{(n)}}{\langle \tilde{F}_c^{(n)} \rangle_c}$$
$$Q_p^{(n)} = \frac{\tilde{Q}_p^{(n)}}{\langle \tilde{Q}_p^{(n)} \rangle_p} \tag{10}$$

The initial conditions are: $\tilde{F}_c^{(0)} = 1$ and $\tilde{Q}_p^{(0)} = 1$; and denominators in the system of Eqs. (10) correspond to the average values for each country and product.

Matrix M_{cp} is usually defined according the Relative Comparative Advantage (RCA) index(Eq. (4)) which determines a relative significance of different products in a country's exports portfolio. However, this index imposes a serious limitation: in this case, countries with high volumes of diversified export products that are uniformly distributed across national exports portfolio, may be ranked lower than countries with a less diversified exports and low export volumes limited to only several items.

To overcome this limitation, we introduce the Relative Effectiveness Advantage (REA) index which defines relative efficiency of specific export items as value of the total exports per capita:

$$REA_{c,p} = \frac{\boldsymbol{X_{c,p}}/\boldsymbol{N_c}}{\sum_c \boldsymbol{X_{c,p}}/\sum_c \boldsymbol{N_c}} \tag{11}$$

Where N_c is the population of the country c.

The two indices are connected in the following way:

$$REA_{c,p} = \frac{g_c}{g} RCA_{c,p} \tag{12}$$

where $g_c = \sum_p X_{c,p}/N_c$ can be considered as the country's effectiveness (i.e. amount of export pec capita), and $g = \sum_{c,p} X_{c,p}/\sum_c N_c$ - as an average effectiveness in the group of countries under consideration.

Results, discussion

Our calculations in respect to the family of three complexity indicators – MECI, ECI, and Fitness, - were performed for a variety of 41 countries, which includes 29 of the 35 OECD

member-states, three BRICS countries (Brazil, China, and Russia), and nine smaller economies: Croatia, Egypt Georgia, Lithuania, Malaysia, Morocco, Moldova, Romania, and Ukraine (*see Appendix, table A1).*
The export products data, given in the format of the Standard International Trade Classification (SITC) revision 3 at the 2-digit level, was retrieved from https://comtrade.un.org.
Data for countries' population and GDP per capita, as well as for the GDP per capita growth were harvested from https://data.worldbank.org.

According to our findings, MECI can support the initial diversity distribution, as shown for illustrative purposes in Fig. 1. In this respect it matches the Fitness index.

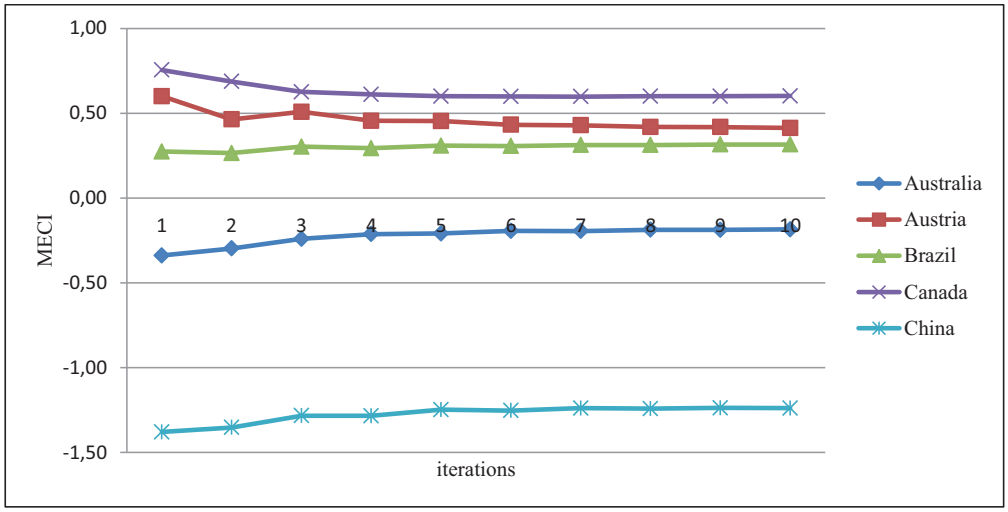


Figure 1: The first ten successive iterations of the Modified Ecosystem Complexity Index for five selected countries (2015 year)

Table 1 shows the Pearson correlations among the values of ECI, Fitness, MECI, initial diversity scores (kc0), nominal GDP per capita, and the logarithm of nominal GDP per capita in current US$ (for 2004). All the three complexity measures were calculated according REA index. One can notice that all the three indices correlate with each other, with MECI and Fitness index correlating more strongly. ECI additionally correlates with the diversity, while MECI and Fitness index highly correlate with LN (GDP per capita), exceeding ECI.in this respect. Whereas in case of RCA index there is no correlation between ECI and initial diversity score, while Fitness strongly correlates with diversity.

Table 1. Pearson correlations (REA-based measures) between the values of MECI, ECI, Fitness, initial diversity score, GDP per capita, and LN (GDP per capita), in current USD, data for 2004

	MECI	*ECI*	*Fitness*	*kc0*	*GDP pc*	*LN (GDP pc)*
MECI	1					
ECI	0,615	1				
Fitness	0,991	0,653	1			
kc0	0,9976	0,599	0,993	1		

GDP pc	0,712	0,560	0,751	0,724	1	
LN(GDP pc)	0,828	0,641	0,844	0,836		1

We also tested ECI, Fitness, and MECI with OLS linear regression growth model for a ten-year time period, by regressing the rate of growth on the initial level of a country's income and complexity index:

$$Growth(t + \Delta t) = A + B \cdot ln(GDP(t)) + C \cdot CI \quad (13)$$

Growth is defined as GDP per capita growth (% for the period), *CI* is a complexity index (correspondingly MECI, ECI, and Fitness index).

Table 2 presents the results of calculations for the three indicators (*t*-values are provided in parentheses) based on the RCA index, while Table 3 presents the same results based on the REA index.
All the three indices demonstrate approximately similar results in regard to R^2 value of the regression, though MECI is slightly better. The Fitness index (with respect to R^2 value) gets improved when moving from RCA index to REA index.

Table 2. OLS ten-year linear regression growth model for RCA based measures

Predicted variable	*Growth (2004, 2014)*	*Growth (2004, 2014)*	*Growth (2004, 2014)*
LN(GDP per capita) - current USD	-51,318 (-6,148)	-52,564 (-7,095)	-54,223 (-7,708)
ECI	-4,239 (-0,398)		
Fitness		-0,371 (-0,234)	
MECI			-5,207 (-0,581)
Constant	578,145 (7,258)	595,969 (9,094)	605,706 (9,001)
Observations	41	41	41
R square	0,615	0,614	0,617

Table 3. OLS ten-year linear regression growth model for REA based measures

Predicted variable	*Growth (2004, 2014)*	*Growth (2004, 2014)*	*Growth (2004, 2014)*
LN(GDP per capita) - current USD	-59,259 (-6,742)	-44,788 (-3,537)	-42,156(-3,503)
ECI	12,014 (1,072)		
Fitness		-0,633 (-0,789)	
MECI			-17,038 (-1,110)
Constant	653,481 (7,796)	535,904 (5,341)	491,242 (4,291)
Observations	41	41	41
R square	0,624	0,620	0,626

The results show that the use of REA instead of the RCA improves the difference among the complexity measurement and introduces refinements in assessing complexity with respect to diversity, the GDP per capita correlation, and the prediction of future growth. In terms of the prediction of growth, MECI provides a slightly better correlation than the other two indicators. The country ranking of competitiveness (Appendix 1) is very similar ($\rho > .99$) for MECI and Fitness though they use different algorithms.

We attribute slightly higher results provided by MECI, as compared to ECI and FCI, to the fact that it takes into account such feature of ecosystems as interactive relationships, considering that in the course of their interactions, an ecosystem's agents and components reflexively adapt to each other's behavior, relying on feedback mechanisms, and that this generates a stream of correlations both in their behavior and in the structure of relationships (Schneider, 2012).

Conclusions

Relying on the CAS theory and considering the ongoing transition of economies to innovation-driven growth, we highlight the ecosystem's approach to measuring complexity. The proposed MECI indicator, as based on this approach, suggests measure the functional complexity of economies in a way that reproduces their organizational complexity. In particular, MECI focuses on incorporating the non-linear and interactive nature of relationships among an economy's agents and components, considering that this enables them to achieve the resulting synergy effects. MECI also considers that the strongest stream of emergent changes (innovations) and the highest yield (critical value of the system) is generated by the maximum degree of diversity of agents (components). Furthermore, MECI takes into account the process of self-organization of ecosystems, implying that a decentralized global order arises from below, as a result of local interactions of many autonomous players (Al-Suwailem, 2011).

Our approach implies that compared to ECI, MECI is improved in terms of revealing a horizontal cohesion of an economy (or vice versa, the scales of its fragmentation), and hence, in reflecting an economy's state of post-industrial transition. In practical terms, this suggests that a more complex (diversified) industrial structure and a higher potential for sustainable growth can be expected in those territories which have developed a better networking context.

While considering non-linear correlations between all estimated parameters, MECI can be easily adjusted to any additional scales, which is not possible for ECI. It can be applied to estimating complexity, and hence, to comparing competitive advantages not just among national or regional economies but also at the level of other types of economic systems, from companies and local clusters to transnational macro-regions. For the same reason, the MECI methodology can be integrated into a range of other economic indicators dealing with estimation of innovativeness and growth potential. Finally, MECI can be extended through incorporating technological complexity data as a third dimension, using, for example, patent portfolios (Ivanova, Strand, Leydesdorff, 2018).

For all these reasons, we believe that the proposed MECI indicator may provide some helpful methodological insights concerned with countries' looking for a better adaptation to unpredictable changes in today's non-linear and globalized world. Noticeably, MECI fits a key idea of the newly emerging patterns of national STI and industrial strategies with the objective to improve the innovation development prospects and positions in globalized markets. Countries should not just follow the optimal directions in exports diversification, as revealed by their ECI indicators, but should also elaborate policy measures towards accelerating the ecosystem-oriented reconstruction of their economic landscapes (Wessner & Wolff, 2012).

Among the two dimensions of relationships in ecosystems, the nonlinear and the interactive one, we have so far formalized only the first — nonlinear interactions. The second one, which gives synergistic effects for endogenous growth, is the topic of future research. The ecosystem's approach leverages the unique role of triple-helix pattern of collaboration (interactive cooperation) among ecosystem's components: both in local and in economy-wide ecosystems such pattern leads to synergies for a continual innovation, and hence, for supporting sustainable growth.

In this paper, we relied on equations tailored to examining complexity in biological ecosystems, which has its own tensions — now we need to proceed with examining the complexity peculiarities (the factor of knowledge flows, etc.) typical particularly for economic ecosystems as specified in the theory about complex adaptive systems.

References

Al-Suwailem, S. (2011). Behavioural Complexity. *Journal of Economic Surveys*, 25(3), 481–506.

Axelrod, R., Cohen, M. D. (2000). *Harnessing Complexity: Organizational Implications of a Scientific Frontier*. New York, NY: Basic Books.

Balassa, B. (1965). Trade Liberalization and "Revealed" Comparative Advantage. *The Manchester School* 33(2), 99–123.

Baldwin, R. (2013). Global Supply Chains: Why They Emerged, Why They Matter, and Where They Are Going. In D. K. Elms, P. Low (Eds.), *Global Value Chains in a Changing World* (pp. 13-59). Geneva: WTO Publications.

Beinhocker, E. D. (2006). *The Origin of Wealth: Evolution, Complexity and the Radical Remaking of Economics*. Boston, MA: Harvard Business School Press.

Hidalgo, C. A., Hausmann R. (2009). The Building Blocks of Economic Complexity. *Proceedings of the National Academy of Sciences of the United States of America,* 106(26), 10570–10575.

Hausmann, R., Hidalgo C. A., *et al*. (2013). The Atlas of Economic Complexity. Retrieved July 15, 2019 from: https://oec.world/static/pdf/atlas/AtlasOfEconomicComplexity_Part_I.pdf

Holland, J.H. (2002). Complex adaptive systems and spontaneous emergence. In: Curzio A.Q., Fortis, M. (Eds.), *Complexity and Industrial Clusters. Contributions to Economics* (pp. 25 –34). Heidelberg: Physica-Verlag.

Ivanova, I., Strand, Ø., & Leydesdorff, L. (2018). An Eco-Systems Approach to constructing Economic Complexity Measures: Endogenization of the technological dimension using Lotka-Volterra equations. *Advances in Complex Systems* (in press).

Leydesdorff, L., Johnson, M., Ivanova, I. (2018). Toward a Calculus of Redundancy: Signification, Codification, and Anticipation in Cultural Evolution. *Journal of the Association for Information Science and Technology* (forthcoming)

Mercan, B., Göktaş D. (2011). Components of Innovation Ecosystems: A Cross-Country Study. *International Research Journal of Finance and Economics*. 76, 102–112.

OECD (2015) Final NAEC Synthesis: New Approaches to Economic Challenges. Paris: OECD Publishing.

Porter, M. E. (1990). *The Competitive Advantage of Nations*. New York, NY: Free Press.

Russell, M.G., Smorodinskaya, N.V. (2018). Leveraging Complexity for Ecosystemic Innovation. *Technological Forecasting and Social Change*. 136, 114-131

Schneider, V. (2012). Governance and Complexity. In D. Levi-Faur (Ed.), *The Oxford Handbook of Governance* (pp. 129–142). Oxford: Oxford University Press.

Tacchella, A., Cristelli, M., Caldarelli, G., Gabrielli, A., & Pietronero, L. (2013). Economic complexity: conceptual grounding of a new metrics for global competitiveness. *Journal of Economic Dynamic and Control,* 37(8), 1683-1691.

Wessner, C.W. & Wolff, A.W. (Eds.). (2012). *Rising to the Challenge. U.S. Innovation Policy for the Global Economy. National Research Council*. Washington (DC): National Academies Press (US).

Yoon J. (2015). The Evolution of South Korea's Innovation System: Moving towards the Triple Helix Model? *Scientometrics,* 104(1), 265–293.

Appendix

Table A1. Comparison of country rankings by MECI, ECI, and Fitness index (2004)

Rank	*country*	*MECI*	*rank*	*country*	*ECI*	*rank*	*country*	*Fitness*
1	Germany	1,32	1	Japan	1,58	1	Germany	60,63
2	Netherlands	1,30	2	China	1,43	2	Switzerland	57,52
3	Switzerland	1,29	3	Korea, Rep,	1,30	3	Netherlands	57,23
4	Ireland	1,24	4	United States	1,22	4	Ireland	56,56
5	Austria	1,20	5	Slovenia	0,73	5	Austria	55,62
6	Sweden	1,18	6	United Kingdom	0,72	6	United Kingdom	55,26
7	Denmark	1,18	7	Luxembourg	0,65	7	France	54,54
8	France	1,16	8	Hungary	0,65	8	Sweden	54,14
9	Canada	1,04	9	France	0,64	9	Canada	53,75
10	United Kingdom	1,01	10	Iceland	0,59	10	Denmark	53,57
11	Finland	1,00	11	Finland	0,59	11	Luxembourg	50,45
12	Spain	0,89	12	Switzerland	0,58	12	Finland	49,97
13	Luxembourg	0,88	13	Ireland	0,54	13	Spain	46,10
14	Czech Republic	0,88	14	Slovak Republic	0,54	14	Czech Republic	45,77
15	Slovenia	0,71	15	Sweden	0,53	15	Slovenia	41,51
16	Slovak Republic	0,60	16	Germany	0,52	16	Norway	40,72
17	Hungary	0,57	17	Austria	0,51	17	Slovak Republic	40,60
18	Norway	0,39	18	Norway	0,50	18	Hungary	39,30
19	New Zealand	0,25	19	Denmark	0,46	19	United States	37,19
20	Portugal	0,14	20	Czech Republic	0,43	20	New Zealand	34,52
21	United States	-0,01	21	Netherlands	0,41	21	Japan	33,15
22	Korea, Rep,	-0,12	22	Spain	0,37	22	Korea, Rep,	32,50
23	Malaysia	-0,16	23	Egypt, Arab Rep,	0,35	23	Malaysia	31,40
24	Japan	-0,17	24	Malaysia	0,32	24	Portugal	29,63
25	Estonia	-0,35	25	Canada	0,29	25	Australia	24,30
26	Australia	-0,42	26	New Zealand	0,17	26	Estonia	23,00
27	Poland	-0,50	27	Portugal	0,01	27	Poland	19,08
28	Lithuania	-0,55	28	Poland	-0,39	28	Iceland	18,19
29	Croatia	-0,61	29	Estonia	-0,52	29	Croatia	16,05
30	Iceland	-0,70	30	Greece	-0,54	30	Lithuania	15,36
31	Latvia	-0,84	31	Australia	-0,59	31	Greece	13,14
32	Greece	-0,87	32	Brazil	-0,59	32	Latvia	9,11
33	Romania	-1,21	33	Croatia	-0,65	33	Russian Federation	6,71
34	Brazil	-1,25	34	Lithuania	-0,72	34	Brazil	6,64
35	Russian Federation	-1,39	35	Romania	-1,07	35	Romania	4,41
36	Morocco	-1,40	36	Morocco	-1,33	36	Morocco	3,22
37	Ukraine	-1,42	37	Latvia	-1,57	37	Ukraine	2,90
38	Georgia	-1,46	38	Moldova	-1,61	38	Moldova	1,66
39	Moldova	-1,50	39	Russian Federation	-1,87	39	Georgia	1,19
40	China	-1,64	40	Georgia	-2,49	40	China	0,00
41	Egypt, Arab Rep,	-1,67	41	Ukraine	-2,70	41	Egypt, Arab Rep,	0,00

Determinants of technology-specific R&D collaboration networks: Evidence from a spatial interaction modelling perspective

Martina Neuländtner[1] and Thomas Scherngell[2]

[1] *martina.neulaendtner@ait.ac.at*
AIT Austrian Institute of Technology GmbH, Giefinggasse 4, 1210 Vienna (Austria)

[2] *thomas.scherngell@ait.ac.at*
AIT Austrian Institute of Technology GmbH, Giefinggasse 4, 1210 Vienna (Austria)

Abstract

It is commonly acknowledged, that the creation of knowledge is the result of interactive, collaborative learning processes among organizations of different types located in different regions. Especially, in a strongly knowledge-based economy built on fast-growing and R&D-intensive technologies such as Key Enabling Technologies (KETs), collaborative knowledge creation is gaining importance to rapidly enable access to external, nation-wide and global new sources of knowledge. With the focus on technology-specific R&D collaboration networks in six KETs, each representing different knowledge bases and modes of (collaborative) knowledge creation, we emphasize the determining role of technology-specific heterogeneities. The objective is to estimate determinants of these technology-specific R&D collaboration networks, focusing on spatial separation and network structural effects. We employ a spatially filtered negative binomial spatial interaction model with a set of 521 regions to identify differences in the determinants of technological knowledge flows, proxied by EU-funded collaborative projects. The results show differences in the relative importance of the determinants. Geographical barriers are significant, and network structural effects are of high importance, but do not remove spatial effects in all KETs. Both spatial and network effects seem to be of higher relevance for more industrial and engineering based than more science based technological fields.

Introduction

Collaborative Research and Development (R&D) activities between firms, universities and research organisations are generally recognized to constitute an essential element for the successful generation of innovation. The notion of *R&D collaboration networks* has come into fairly wide use for describing such collaborative endeavours (see Barber and Scherngell 2013) and has become a major research domain in manifold aspects. One major research stream is – without doubt – the identification and estimation of determinants affecting structures and dynamics of such networks, often with a geographical focus and accomplished at the regional level of analysis (see Scherngell and Barber 2009; Hoekman et al. 2010; Scherngell and Lata 2013; Lata et al. 2015). However, these works capture R&D, and accordingly the underlying knowledge, in a quite aggregated manner, neglecting technology-specific peculiarities of knowledge creation and interactions, such as technological regimes, as well as different modes of (collaborative) knowledge creation.

This study intends to address this research gap by accounting for technological idiosyncrasies when explaining the constitution and dynamics of R&D collaboration networks. Accordingly, the objective is to estimate determinants of technology-specific R&D collaboration networks, shifting particular attention – as in previous works – to spatial separation effects, such as geographical distance or country borders, but also to network structural effects, such as central positioning, influencing the collaboration probability between two organisations. To address this objective, we employ a spatial interaction modelling approach at the regional level of analysis. The R&D collaboration network under consideration is the project-based network of organisations that collaborate in projects funded by the EU Framework Programme (FP). This network is partitioned into different technological domains and aggregated from the organisational to the regional level of analysis, using a set of 521 European metropolitan and remaining non-metropolitan regions. The technological disaggregation is attained by assigning collaborative projects to specific relevant technologies. In the latter context, we use the so-

called Key Enabling Technologies (KETs), considered by the EU as specifically relevant in the global innovation competition. Semantic technologies are used to assign projects to KETs based on sets of keywords and semantic characteristics.
In what follows, we discuss in some more detail the relevance of technological heterogeneities in the context of R&D collaboration, before we introduce the model, the data, and variables. The study closes with a compact presentation and discussion of the results, and some ideas for a future research agenda.

Technological heterogeneities in R&D collaboration networks

Previous studies that identify determinants of R&D collaboration networks, find quite robust and increasingly stylized results, in particular in terms of their spatial dynamics. Spatial proximity turns out to be an important factor for the constitution of R&D collaboration, also in times of increasing globalisation and new communication technologies (see e.g. Scherngell and Barber 2009). However, while geographical barriers seem still to be significant, they tend to decrease in terms of their relative importance as when compared to other forms of separation, such as technological distance (see e.g. Scherngell and Lata 2013), and/or network structural effects (see Autant-Bernard et al. 2007; Broekel 2012).
Although these insights are interesting and have substantially increased the understanding on structure and dynamics of such networks, the main limitation is the disregard of technological heterogeneities that may influence the relevance and spatial scale of R&D collaboration (see Ponds et al. 2007; Martin and Moodysson 2013). Therefore, this study intends to shift the focus on the debate of the differing role of main determinants for R&D collaboration networks in different technologies. Concerning technological differences, especially novel and fast-growing technologies that spur innovation and technological progress of countries, regions and industries have gained anew interest. At the European policy level, this is reflected by the new emphasis on so called Key Enabling Technologies (KETs) bringing technologies into focus that are considered as crucial for the development of the EU towards a sustainable, knowledge-based economy (EC 2009).
KETs are understood as generic technologies, which are characterised by relatively rapid pervasiveness and growth. They constitute technological inputs for the development of innovation, and by this require high R&D intensity and a high input of skilled labour in their creation. The European Commission defined six KETs: N*anotechnology*, *Micro- and Nanoelectronics*, *Photonics*, *Advanced materials (AM)*, *Advanced manufacturing technology (AMT)* and *Industrial biotechnology* (EC 2009). Due to the specific characteristics of KETs – knowledge intensive, high R&D intensity, rapid innovation cycles, highly skilled employment etc. (EC 2009) – R&D networks are considered of particular importance in order to cope with the high demand for R&D and to gain rapid access to nation-wide and global state of the art knowledge. Specifically, in such globally relevant technologies, R&D networks may serve as channels for transmitting knowledge over larger geographical distances (see e.g. Autant-Bernard et al. 2007), and hence be of particular importance for innovation and regional growth processes (Huggins and Thompson 2014). Hence, stimulating knowledge creation and interaction in KET fields has become one of the major priorities of the EU industrial policy to accelerate industrial restructuring and change, particularly in structurally weak regions.
With our focus on networks of KETs, we propose – in contrast to previous research – a finer grained and policy relevant perspective when identifying determinants of R&D collaboration networks. Collaborative research activities follow different rationales and aims – especially across different technologies, leading to different outcomes of knowledge creation, that have not been accounted for so far in the literature; comprehensive investigations and studies allowing for comparisons on the role of networks in different technological fields are still missing. Hence, we go beyond the beaten track by investigating determinants of technology-specific R&D collaboration networks – proxied by KET fields – where we especially focus on

spatial and network structural effects. We employ a comparative perspective on the effect sizes of the determinants of KET collaboration networks to gain insights in technology-specific heterogeneities.

Satisfying the multidisciplinary character of KETs, cutting across many technological domains on regional, national and subnational levels, makes a spatial network perspective with a spatial interaction modelling approach evident. By this, this research will be the first, bridging networks and KETs from a spatial perspective to examine systematically the (claimed) converging and integrating nature of KETs, explicitly considering the systemic, cross-sectoral and inter-regional character of KETs.

Methodological approach and model

For the estimation of spatial and network structural determinants on technology-specific R&D collaboration networks, we follow earlier research and employ a spatial interaction modelling approach. Spatial interaction models refer to a class of models applied to identify determinants – particularly separation effects – of interactions between discrete spatial entities (Roy and Thill 2003), such as in our case interactions in R&D collaboration networks between regions. In general, these types of models comprise three types of factors to explain mean interaction frequencies between spatial locations i and j. The general form of the model can be written as

$$Y_{ij} = \mu_{ij} + \varepsilon_{ij} \quad \text{with} \quad i, j = 1, \dots, N \tag{1}$$

where $\mu_{ij} = E(Y_{ij})$ is the expected mean interaction frequency from i to j, and ε_{ij} is an error about the mean (Fischer and Wang 2011).

In this specific context of application, locations correspond to European metropolitan and remaining non-metropolitan regions, where each location is both origin and destination of interactions. Accordingly, the model class distinguishes: (i) *origin-specific* factors characterising the ability of the origins to generate R&D network links, (ii) *destination-specific* factors indicating the attractiveness of destinations, and (iii) *separation factors* that represent the way different forms of *separation* between origins and destinations constrains or impedes the interaction, most basically geographical distance (LeSage and Fischer 2016). Hence, the mean interaction frequencies between origin i and destination j are modelled by

$$\mu_{ij} = O_i \, D_j \, S_{ij} \quad \text{with} \quad i, j = 1, \dots, N \tag{2}$$

where O_i and D_j are the origin-specific and destination-specific factors, respectively, and S_{ij} denotes a multivariate function of separation between locations i and j.

While there are different functional forms used to specify origin-, destination- and separation functions (see Fischer and Wang 2011), studies investigating R&D networks usually employ univariate (i.e. with only variable) power functional forms for origin and destination functions, and multivariate (i.e. with a number of separation variables) exponential functional forms for the separation function. We follow these lines and define

$$O_i = O(o_i, \alpha_1) = o_i^{\alpha_1} \tag{3}$$

$$D_j = D(d_j, \alpha_2) = d_j^{\alpha_2} \tag{4}$$

$$S_{ij} = \exp\left[\sum_{k=1}^{K} \beta_k s_{ij}^{(k)}\right] \tag{5}$$

Here, o_i and d_j are measured in terms variables controlling for the mass in the origin and the destination, respectively. In context of R&D networks, these are often captured by the number of firms or researching organisations in a region. α_1 and α_2 are scalar parameters to be estimated, so that the product of the functions $O_i D_j$ can be simply interpreted as the number of

cross-region R&D collaborations which are possible. Core of the spatial interaction model is the separation function as defined by Equation (5), with K (k = 1, …, K) separation measures to be estimated that will show the relative strengths of the separation measures, and β_k denoting the respective k^{th} estimate for separation measure k.

The model applied here takes the specific form of a spatially filtered, negative binomial spatial interaction model (see Scherngell and Lata 2013 in a similar context)[1]. The main motivation for this is given by the true integer nature and distributional assumptions on the dependent variable, cross-region R&D collaborations. Further, the proposed model specification accounts for the spatial dependence of the data used (participation in European Framework Programme (FP) projects) in the empirical application, as well as for a high degree of variation (overdispersion) and a large amount of zero counts. Hence, it is assumed that the dependent variable Y_{ij} follows a negative binomial distribution with expected values as stated in (2).

In comparison to the Poisson model that assumes equidispersion (i.e. conditional mean equals the conditional variance), the negative binomial model explicitly corrects for overdispersion[2], by adding a dispersion parameter γ. Hence, the negative binomial spatial interaction model takes the form

$$\Pr(Y_{ij} = y_{ij}|\mu_{ij}, \gamma) = \frac{\Gamma(y_{ij} + \gamma^{-1})}{\Gamma(y_{ij} + 1)\Gamma(\gamma^{-1})}\left(\frac{\gamma^{-1}}{\gamma^{-1} + \mu_{ij}}\right)^{\gamma^{-1}}\left(\frac{\mu_{ij}}{\gamma^{-1} + \mu_{ij}}\right)^{y_{ij}} \tag{6}$$

where $\mu_{ij} = E[y_{ij}|O_i, D_j, S_{ij}] = \exp[O_i(\alpha_1)\, D_j(\alpha_2)\, S_{ij}(\beta)]$ and Γ denotes the gamma function with a model parameter γ accounting for overdispersion in predictors (see Cameron and Trivedi 1998 for a more detailed derivation).

To take the spatial dependence of flows into account, spatial filtering using eigenvectors (ESF) is employed[3]. ESF is based on the mathematical relationship between the Moran's I, as a measure for spatial autocorrelation, and spatial weights matrices. Following Griffith and Chun (2014), the purpose is to obtain a set of synthetic proxy variables by extracting them as eigenvectors from a standard spatial weights matrix (see e.g. Fischer and Wang (2011) on construction of spatial weights matrices), and then add these vectors as control variables to the regression model.

In this study, six separate – one for each KET – regression models are estimated via the spatially filtered negative binomial spatial interaction model. We include the first ten eigenvectors from the set κ of eigenvectors with MI/MI_{max} larger than 0.25, where MI denotes the Moran's I value and MI_{max} its maximum value, as additional explanatory variables in the model (see e.g. Fischer and Wang (2011) for details).

Recalling the negative binomial specification of the model in (6), the full empirical model is specified by setting

$$\mu_{ij} = \exp(\alpha_0 + \alpha_1 \ln(o_i) + \alpha_2 \ln(d_j) + \sum_{k=1}^{K} \beta_k s_{ij}^{(k)} + \sum_{q=1}^{Q} \phi_q E_q + \sum_{r=1}^{R} \varphi_r E_r + \xi_{ij} \tag{7}$$

where E_q denotes the selected subset of eigenvectors expanded by means of the Kronecker product associated to the origin variable, and E_r the respective eigenvectors for the destination

[1] Although the data used has excess zeroes, we did not opt for a zero-inflated version of the negative binomial model, since we argue that each region possibly has the chance to engage in a collaboration (no structural zeroes).

[2] Not accounting for overdispersion would result in incorrect standard errors, leading to possibly wrong significances of parameters (Cameron and Trivedi 1998).

[3] In the context of spatial interactions, spatial autocorrelation of flows is understood as correlation between R&D collaboration flows from the same origin or destination, to neighboring origins or destinations, respectively. Not accounting for spatial autocorrelation leads, similar to overdispersion, to incorrect inferences and hence, wrong significances (Chun 2008).

variable; ϕ_q and φ_r are the corresponding coefficients. Note, that the explanatory variables enter the regression in their logged form (except the dummy variables).
Since the assumption of the dependent variable – the R&D interactions between region i and j – being independent and normally distributed does not hold, the parameters of the model are estimated by means of Maximum Likelihood (ML) estimation.

Data and variables

The main interest of this study is to estimate determinants of technology-specific R&D collaboration networks, with a special focus on spatial separation and network structural effects. The geographical coverage comprises the currently 28 EU member states, plus Switzerland and Norway, corresponding to a set of 521 regions. Going beyond previous research, we distinguish metropolitan regions as well as remaining non-metropolitan regions based on the 2013 NUTS version and the 2010 Geostat population grid defined by Eurostat[4].

Dependent variable

As dependent variable EU-funded KET R&D collaboration links are used (see Table 1 for some descriptive statistics). Data is extracted from the EUPRO data base comprising systematic information on collaborative research projects of FP1–FP7 as well as Horizon 2020 (until 2016), including information on respective participating organizations, e.g. name and type or participating organization and their geographical location in the form of organization addresses (see Heller-Schuh et al. 2015 for details). The latter is used to geolocalize participating organizations and to assign them to regions, enabling the observation of region-level R&D collaboration activities[5].

Table 1. Descriptive statistics on R&D collaborations in six KETs

	Nano	Micro	Photonics	AM	AMT	Biotech
# All links	271441	271441	271441	271441	271441	271441
# Positive links	18434	6923	19229	2641	9169	32867
% Zero links	93.21	97.45	92.916	99.03	96.62	87.89
# Intra-regional collaborations	909	358	1147	44	266	2251
# Inter-regional collaborations	48852	15500	54052	3634	17246	129634
# Organizations	2221	850	2391	305	1035	3911

Notes: Nano = Nanotechnology, Micro = Micro- and Nanoelectronics, AM = Advanced materials, AMT = Advanced manufacturing technologies, Ind. Biotech. = Industrial biotechnology

To construct the dependent variable, we consider the 7th FP and H2020, i.e. a time horizon of 2007-2016. For each KET a technology-specific symmetric regional collaboration matrix is constructed, where the elements indicate the number of joint EU-funded research projects[6]. This matrix is then transformed into a vector with rows representing all possible combinations of links between the regions; this results in a vector of length n^2-by-1 containing the inter- and intra-regional collaboration activities of all region pairs.

[4] Although the NUTS-2 level perspective is widely used in previous related empirical literature (e.g. Scherngell and Barber 2009; Hoekman et al. 2012), we opt for metropolitan regions as units of analysis. Metropolitan regions are a quite recently introduced classification on a European level based on agglomeration (EC 2008; Dijkstra 2009), which by definition is an urban core including the surrounding catchment area. Hence, this classification corrects for distortions created by e.g. the NUTS classification that separates these two geographical spaces in most cases.

[5] The EUPRO database in maintained by AIT and is accessible via RISIS (risis2.eu). It has been advanced within RISIS, in particular in terms of geolocalisation, standardization and integration with other datasets.

[6] The number of collaborations between regions results from the aggregate of collaborations (full count) between the participating organisations located within these regions.

Independent variables

As described in the previous section, the independent variables comprise three types of variables: origin-, destination- and separation variables. The origin variable o_i and the destination variable d_j are solely specified as the number of organizations participating in joint EU-funded FP projects in region i and j in a distinct KET field. For the separation variables, we distinguish between (i) spatial separation variables, (ii) network structural separation variables, as well as (iii) a measure for technological separation. Empirically, these variables represent the potential of regions to engage in collaborative R&D activities. Statistically, they control for the different sizes of the regions.

The separation variables constitute the core of the spatial interaction model since they can be interpreted as determinants of the technology-specific R&D collaboration networks. The focus of this study lies on the spatial as well as network structural determinants:

- As variables accounting for spatial separation effects, *first*, the geographical distance, measured as the great circle distance, indicating the shortest distance between two regions i and j, *second*, a dummy variable indicating the presence of a common national border of regions (set to one, if two regions are located in different countries, zero otherwise), and *third*, a dummy variable indicating links between two metropolitan regions (set to one, if link between two metropolitan regions, zero otherwise), are included in the model.
- As network structural separation effects, *first*, the gap in degree centralities and *second*, the gap in the hub score between the two regions i and j, are included. Whereas, the degree centrality simply measures the number of collaboration links of a region, the hub score (Kleinberg's authority centrality[7]) is defined as the principal eigenvector of $A\,t(A)$, where A is the adjacency matrix of the KET-specific R&D network and hence, indicates whether a region holds reliable information on the topic of interest and at the same time is linked to other regions, themselves with reliable information. Together, the two variables account for differences in the quantity of collaboration links, as well as difference in the quality of these interactions.
- As a last separation variable, the technological distance is included. It is constructed by using patent data drawn from the IFRIS-PATSTAT, which provides KET-specific structured information on patent applications including details on the patent itself, e.g. date of application and technology classes, as well as information on applicants and inventors, such as their names and location[8]. Technological distance between two regions i and j is defined as the correlation between the vectors of patent applications in 353 KET-subtopics (see '*Assignment of data items to KETs'* for details); the technological distances between regions is KET-specific.

Assignment of data items to KETs

The meaningful delimitation of KETs in EUPRO and IFRIS-PATSTAT is essential to address the research objectives of this study. However, KETs are usually cross-cutting technological domains, and are not pre-defined categories in a both datasets under consideration. Thus, we employ the classification approach developed in the EU-funded project KNOWMAK that provides an ontology for KETs, comprising a hierarchical system of topical classes for each KET that are characterised by a set of weighted keywords. The data items are assigned to these topical classes, based on a scoring system that evaluates the similarity of a text (in our case an abstract of a FP project or patent) to a keyword set of a specific KET-related subtopic. By this, projects and patents are tagged to specific KET subtopics which are aggregated to the six main

[7] Equals the authority score for undirected graphs.

[8] IFRIS-PATSTAT is based on the PATSTAT database (developed by the European patent office) and accessible via RISIS (risis.eu). It shifts attention to organisation names cleaning of PATSTAT, geolocalisation, and most importantly in our context, assignment of patents to KETs.

KETs to extract the six KET-specific collaboration networks and patenting activities for the analysis at hand. Note that assignment of projects is subject to a series of robustness and sensitivity analysis (including manual checking of individual cases) to guarantee a sufficiently meaningful and robust result (see also Maynard et al. 2017)[9].

Estimation results

In Table 2, the estimation results of the spatial interaction models are displayed. While the first column reports the ML estimates for a basic spatial interaction model (model 1), including the origin and destination variables as well as the geographical distance and the country border effect as separation measures, the second column comprises the results for the full model (model 2) including an additional set of spatial and network structural separation measures. Each of the two model specifications was executed for all six KETs to allow the comparison between the effect sizes of the determinants of technology-specific R&D collaboration networks. For all models, the significance of the γ-parameter suggests the preference of a negative binomial model over the Poisson specification. Moreover, a likelihood ratio test was conducted testing the spatially filtered negative binomial model against the non-filtered version, clearly pointing towards the filtered adaptation for all models. Note that we aggregate over the whole time period (i.e. summing up FP7 and H2020 due to the extremely high number of zeros challenging a reasonable estimation.

In our discussion, we focus on the separation variables. As can be seen from Table 2, the origin- and destination variables that just control for the mass in the origin and the destination region are significant and higher than one, i.e. the number of organisations active in a KET in a region obviously increases the likelihood for R&D collaboration in this KET with other regions. Turning to the results of the separation effects for model (1), it can be seen that the geographical distance between two regions has negative effect on the expected collaboration frequency between these two regions for all KETs – as indicated by the negative and significant estimates; this result coincides with findings in previous studies (Scherngell and Barber 2009; Scherngell and Lata 2013). Whereas, the effects are the highest (the most negative) for Industrial biotechnology (for a coefficient of -0.277 this equals to a change of -0.24% given by its exponential[10], followed by Nanotechnology (with a factor change of 0.78; i.e. a change of - 0.22%), the effect for Advanced materials (AM) is the smallest with a factor change of 0.90 (i.e. a change of -0.10%).

The coefficients for the country border effects are also significantly negative for all KETs, suggesting that a national border between any two regions decreases the expected collaboration frequency for participating organisations located in these regions. This is a rather negative outcome in a European integration and policy context. While country border effects seem to diminish in networks of the FP as a whole as evidenced by Scherngell and Lata (2013), in KETs – that are considered as the most important technological domains for economic competitiveness – they are still a significant barrier for collaboration. Interestingly, here the negative effects are the lowest for Nanotechnology and Industrial biotechnology, i.e. the more science-based fields, while AM shows by far the highest negative effect. For region pairs located in different countries the expected number of collaborations is hypothetically decreased by 49% in the case of AM.

Model (2) adds the technological distance and network structural separation variables, as well as the metropolitan region dummy as additional spatial separation. Interestingly, the interpretation of the coefficient for the variables already included in model (1) stays the same,

[9] Details on the semantic approach and also the technical tools are given at knowmak.eu

[10] A change of one kilometre in geographical distance results in an expected count decrease by a factor of $\exp(-0.277) = 0.758$ which implies a change of -0.24% see Long and Freese (2006).

Table 2. Estimation results of the spatially filtered negative binomial spatial interaction models

	Model (1)						Model (2)					
	Nano	Micro	Photonics	AM	AMT	Ind. Biotech.	Nano	Micro	Photonics	AM	AMT	Ind. Biotech.
Origin and destination variable [$\alpha_1 = \alpha_2$]	1.411*** (0.007)	1.756*** (0.014)	1.427*** (0.008)	2.631*** (0.035)	1.775*** (0.014)	1.389*** (0.005)	1.458*** (0.009)	1.963*** (0.017)	1.482*** (0.009)	2.940*** (0.041)	1.910*** (0.016)	1.441*** (0.006)
Geographical distance [β_1]	-0.243*** (0.011)	-0.235*** (0.017)	-0.208*** (0.011)	-0.107*** (0.028)	-0.210*** (0.016)	-0.277*** (0.008)	-0.196*** (0.011)	-0.162*** (0.018)	-0.150*** (0.011)	-0.018 (0.029)	-0.173*** (0.016)	-0.233*** (0.008)
Country border effects [β_2]	-0.089*** (0.030)	-0.126** (0.050)	-0.131*** (0.029)	-0.398*** (0.082)	-0.189*** (0.044)	-0.094*** (0.022)	-0.118*** (0.030)	-0.235*** (0.052)	-0.186*** (0.029)	-0.503*** (0.086)	-0.225*** (0.044)	-0.124*** (0.021)
Technological distance [β_3]	-	-	-	-	-	-	-0.352*** (0.069)	-0.065 (0.120)	-0.358*** (0.070)	0.008 (0.198)	-0.184 (0.105)	-0.295*** (0.050)
Metropolitan region [β_4]	-	-	-	-	-	-	0.229*** (0.016)	0.238*** (0.027)	0.117*** (0.016)	-0.064 (0.045)	0.000 (0.024)	0.182*** (0.011)
Gap in degree centralities [β_5]	-	-	-	-	-	-	-0.128*** (0.009)	-0.277*** (0.014)	-0.166*** (0.008)	-0.440*** (0.027)	-0.128*** (0.013)	-0.148*** (0.006)
Gap in hub score [β_6]	-	-	-	-	-	-	-0.885*** (0.077)	-1.574*** (0.120)	-0.562*** (0.077)	-0.079 (0.194)	-1.518*** (0.116)	-0.963*** (0.057)
Constant [α_0]	-5.467*** (0.078)	-5.962*** (0.127)	-5.829*** (0.080)	-7.848*** (0.209)	-6.298*** (0.115)	-5.492*** (0.059)	-5.407*** (0.080)	-6.090*** (0.127)	-5.723*** (0.081)	-7.835*** (0.214)	-6.328*** (0.116)	-5.341*** (0.060)
Dispersion [γ]	1.052 (0.020)	1.712 (0.019)	1.303 (0.015)	2.070 (0.027)	1.770 (0.016)	0.922 (0.016)	1.038 (0.020)	1.739 (0.017)	1.291 (0.014)	2.503 (0.020)	1.750 (0.015)	0.860 (0.017)
Likelihood ratio test	924.4***	438.5***	601.7***	356.4***	564.1***	1628.5***	832.6***	370.3***	517.1***	265.9***	497.1***	1479.9***

Notes: The dependent variable is the number of EU-funded R&D collaborations between two regions; each ten origin and destination spatial filters as specified in the text are included as explanatory variables; the number of observations is 271441; standard errors are given in parentheses; *** indicates significance at the 0.001 level, ** indicates significance at the 0.01 level, *indicates significance at the 0.05 level; due to the symmetry of the origin and destination variable, α_1 equals α_2 up to numerical precision; the Likelihood ratio test compares tests the spatial filtered model against the non-filtered equivalent (Chi-squared with 20 degrees of freedom); Nano = Nanotechnology, Micro = Micro- and Nanoelectronics, AM = Advanced materials, AMT = Advanced manufacturing technologies, Ind. Biotech. = Industrial biotechnology

in terms of significance and direction; however, geographical distance decreases in magnitude when adding the additional variables, i.e. geographical distance may partly be a proxy for the other effects reflected in the additional variables, in particular the metropolitan dummy. The latter indicates links between two metropolitan regions. The estimate is positive and significant for the KETs Nanotechnology, Microelectronics, Photonics and Industrial Biotechnology. This implies that two metropolitan regions 'increase' the expected number of collaborations of their organizations by +12% in the field of Photonics, that exhibits the smallest effect, and +27% in Microelectronics with the largest effect (compared to links between non-metropolitan regions and links between metropolitan and non-metropolitan regions).

Turning to technological distance that accounts for technological effects that may determine the collaborative activities of organisations located in the regions of interest, only for the KETs Nanotechnology, Photonics and Industrial Biotechnology the coefficient is of significance in explaining inter-regional R&D collaborations. Especially, these KETs are characterised in the literature as cross-sectoral technologies that combine approaches from physics, chemistry and biology, materials science and electrical engineering (Aschhoff et al. 2010). Comprising many scientific fields, and hence being quite heterogeneous in their technological focus suggests possibly large technological distances that have hampering effects on the collaborative activities between regions – as evidenced by the significantly negative estimates for these KETs.

The addition of network structural effects is another element where this study goes beyond previous research. The results are highly interesting, *first*, the spatial effects remain largely unchanged, showing that geographical variables are not simply proxies of underlying network structural effects, and *second*, that such network structural effects are important. We find a significantly negative impact of the gap in degree centralities between two regions on their expected collaboration frequency – in all KETs. That is, the number of collaborations is expected to be higher between similar regions in terms of the *quantity* of existing collaboration links. The effects of the gap in hub score point towards the same direction, being negative and significant for all KETs except AM. However, the gap in hub score can be interpreted as the gap in *quality* of the region in providing knowledge and enabling knowledge access. In other words, collaboration probability between two regions decreases when their difference in terms of quantity and quality of links increases, i.e. hubs are more likely to connect with other hubs than to connect with peripheral regions, which is described as preferential attachment mechanisms from a network perspective.

KET-specific differences seem to be of minor relevance in terms of the gap in degree centralities, i.e. quantity of the links. However, for the gap in the hub score, i.e. the quality of links, we find some notable differences between the KETs. In Micro- and Nanoelectronics and Advanced manufacturing technology (AMT) hub score effects are by far highest, suggesting a distinguished authorities- and hub-structured network for these KETs.

Concluding remarks

The investigation of structures and dynamics of R&D collaboration networks has become one of the most important research domains in Science, Technology and Innovation (STI) studies, accounting for their essential influence for successfully generating new knowledge, and accordingly, innovation. In the recent past, attention has been shifted to get more comprehensive and statistically robust insights into R&D network dynamics by systematically identifying and estimating determinants and drivers of real-world observed network structures. The number of empirical works embedded in this research vein has faced an upsurge over the past ten years, related to methodological advancements, but more importantly to the recent establishment of large-scale databases enabling to trace R&D collaboration networks in space and time, covering increasingly large geographical areas and time periods (see, e.g. risis.eu).

Empirical studies investigating determinants of R&D collaboration networks – mostly done at the regional level of analysis – have so far brought highly interesting results (see Scherngell 2019 for an overview), pointing to the still important role of geographical barriers (geographical distance and/or country borders) and technological determinants, such as technological distance. However, the studies so far did not dig yet into technological differences that may be prevalent across these results. Such technological heterogeneities are assumed to play a major role, given the different knowledge bases and knowledge creation regimes underlying different technological fields, and accordingly different collaboration behaviours.

This study has directly addressed this research gap, aiming to identify determinants of technology-specific R&D collaboration networks across a set of European regions. We have employed a spatially filtered negative binomial spatial interaction model to estimate a set of determinants, specifically focusing on spatial effects, and – in contrast to previous works – on network structural effects. By technology-specific networks, we refer to collaborative R&D projects of the EU Framework Programme (FP) observed in six Key Enabling Technologies (KETs), giving rise to six cross-region European R&D networks in different relevant technologies. In our empirical strategy, we have used the EUPRO database on EU-FP projects, that contains an assignment of projects to a specific KET based on semantic technologies (see Maynard et al. 2017). The spatial interaction models are applied to each KET separately and aggregated for FP7 and H2020 for a system of 521 European metropolitan and remaining non-metropolitan regions, relating the cross-region collaboration intensity to a set of exogenous variables, in particular, spatial and network structural separation variables.

The results are highly interesting, both in context of previous research and from a European policy perspective. *First*, geographical barriers, including geographical distance and country borders are a significant hurdle for the likelihood to establish network links across regions in the six KETs. While the negative effect of geographical distance is not surprising, and also not tremendously high, the significant country border effects are somewhat negative in a policy context. Negative country border effects have diminished when looking at the FP as a whole (see Scherngell and Lata 2013) but are back at stake when looking at important technological fields, such as the KETs. *Second*, network structural effects turned out to be indeed an important additional determinant that has been neglected in previous works, in particular, pointing to the existence of preferential attachment mechanisms, i.e. regions of similar network embeddedness are more likely to collaborate than regions with a high gap in network embeddedness. Accordingly, lagging regions in terms of network centrality face statistically significant barriers to attach to more prominent regions in the network. *Third*, we find indeed significant and very relevant differences between the KETs under consideration, not in terms of direction and significance of the effects, but in terms of their relative importance. The more science-based KETs (Nanotechnology and Biotechnology) seem to be less affected by geographical barriers than the more engineering and industrial driven fields (Advanced materials and Advanced manufacturing technology). For the latter, network structural effects seem to be of relatively

higher importance, i.e. science-based fields may be more open to non-conventional network partners than industry driven fields.

Some ideas for a future research agenda come to mind. *First*, the semantic approach to assign R&D projects to technologies is based on a first version as described in Maynard et al. (2017). Updated versions of the ontology may alter the results. This needs to be checked in terms of robustness of the results. *Second*, the results presented in this study are static, mainly relating to the problem of the high number of zeros when going to a panel with annual observations, leading to severe estimation issues. However, advancement to a dynamic perspective to look at changes of the estimates over time is crucial and needs consideration in the future. *Third*, looking at other forms of technology-specific R&D networks should complement the results of this study that clearly focuses on a specific form of policy induced networks.

Acknowledgements

This research greatly benefited from the data base infrastructure built up in the two EU-funded projects RISIS (risis2.eu) and KNOWMAK (knowmak.eu).

References

Aschhoff, B., D. Crass, K. Cremers, C. Grimpe, C. Rammer, F. Brandes, F. Diaz-Lopez, R. K. Woolthuis, M. Mayer and C. Montalvo (2010). European competitiveness in key enabling technologies. *Background Report to the EU Competitiveness Report*.

Autant-Bernard, C., J. Mairesse and N. Massard (2007). Spatial knowledge diffusion through collaborative networks. *Papers in Regional Science* 86(3): 341-350.

Barber, M. J. and T. Scherngell (2013). Is the European R&D network homogeneous? Distinguishing relevant network communities using graph theoretic and spatial interaction modelling approaches. *Regional Studies* 47(8): 1283-1298.

Broekel, T. (2012). Collaboration intensity and regional innovation efficiency in Germany—a conditional efficiency approach. *Industry and Innovation* 19(2): 155-179.

Cameron, C. and P. Trivedi (1998). Models for count data, Cambridge University Press, Cambridge.

Chun, Y. (2008). Modeling network autocorrelation within migration flows by eigenvector spatial filtering. *Journal of geographical systems* 10(4): 317-344.

EC (2009). Preparing for our future: Developing a common strategy for key enabling technologies in the EU. Brussels, European Commission.

Fischer, M. M. and J. Wang (2011). *Spatial data analysis: models, methods and techniques*. Springer Science & Business Media, Berlin.

Griffith, D. and Y. Chun (2014). Spatial autocorrelation and spatial filtering. In: Fischer, M. M. and P. Nijkamp (eds) *Handbook of Regional Science* Springer, Berlin Heidelberg: 1477-1507.

Heller-Schuh, B., M. Barber, M. Züger and T. Scherngell (2015). Report on the content and technical structure of the EUPRO infrastructure, AIT Austrian Insitute of Technology.

Hoekman, J., K. Frenken and R. J. Tijssen (2010). Research collaboration at a distance: Changing spatial patterns of scientific collaboration within Europe. *Research Policy* 39(5): 662-673.

Huggins, R. and P. Thompson (2014). A network-based view of regional growth. *Journal of Economic Geography* 14(3): 511-545.

Lata, R., T. Scherngell and T. Brenner (2015). Integration Processes in European Research and Development: A Comparative Spatial Interaction Approach Using Project Based Research and Development Networks, Co-Patent Networks and Co-Publication Networks. *Geographical Analysis* 47(4): 349-375.

LeSage, J. P. and M. M. Fischer (2016). Spatial regression-based model specifications for exogenous and endogenous spatial interaction. In: Patuelli, R. and G. Arbia (eds) *Spatial Econometric Interaction Modelling* Springer International Publishing, Switzerland: 15-36.

Long, S. J. and J. Freese (2006). *Regression models for categorical dependent variables using Stata*. Stata press.

Martin, R. and J. Moodysson (2013). Comparing knowledge bases: on the geography and organization of knowledge sourcing in the regional innovation system of Scania, Sweden. *European Urban and Regional Studies* 20(2): 170-187.

Maynard, D., A. Funk and B. Lepori (2017). Towards an Infrastructure for Understanding and Interlinking Knowledge Co-Creation in European Research. Lecture Notes in Computer Science (Proceedings of ESWC 2017 Workshop on Scientometrics), Springer Verlag.

Ponds, R., F. Van Oort and K. Frenken (2007). The geographical and institutional proximity of research collaboration. *Papers in Regional Science* 86(3): 423-443.

Roy, J. R. and J.-C. Thill (2003). Spatial interaction modelling. *Papers in Regional Science* 83(1): 339-361.

Scherngell, T. (2019). The Geography of Networks and R&D Collaboration. A review. *AIT working paper* 01/2019.

Scherngell, T. and M. J. Barber (2009). Spatial interaction modelling of cross-region R&D collaborations: Empirical evidence from the 5th EU framework programme. *Papers in Regional Science* 88(3): 531-546.

Scherngell, T. and R. Lata (2013). Towards an integrated European Research Area? Findings from Eigenvector spatially filtered spatial interaction models using European Framework Programme data. *Papers in Regional Science* 92(3): 555-577.

Do national funding organizations address the diseases with the highest burden adequately? ---- Observations from the China and UK

Lin Zhang[1*], Wenjing Zhao[2], Jianhua Liu[3], Gunnar Sivertsen[4], Ying Huang[5*]

[1*] *zhanglin_1117@126.comn*
School of Information Management, Wuhan University, Wuhan 430072 (China)
Department MSI, Centre for R&D Monitoring (ECOOM), KU Leuven, Leuven B-3000 (Belgium)

[2] *cady_zhao0827@163.com*
School of Information Management, Wuhan University, Wuhan 430072 (China)

[3] *liujh@shanghaitech.edu.cn*
Library and Information Services, ShanghaiTech University, Shanghai 201210 (China)

[4] *gunnar.sivertsen@nifu.no*
Nordic Institute for Studies in Innovation, Research and Education (NIFU), Oslo (Norway)

[5*] *huangying_work@126.com*
School of Information Management, Wuhan University, Wuhan 430072 (China)
Department MSI, Centre for R&D Monitoring (ECOOM), KU Leuven, Leuven B-3000 (Belgium)

Abstract

This paper is aimed to explore whether health research funding organizations have paid enough attention to the diseases with the highest burden. We studied the evolution of the hottest research topics through projects funded by the Department of Health Sciences, National Natural Science Foundation of China (NSFC) and the Medical Research Council (MRC) of the UK during 2006-2017, and compared the focus of these funded research projects against the diseases that carry the highest burden. The results indicate that both the NSFC and the MRC are greatly concerned with neoplasms and cardiovascular diseases, which do correspond to the top two families of diseases with the highest burden. Another family of diseases of common concern is diabetes and kidney diseases, which have shown an increasing trend in both burden and research attention since 2006. The MRC has funded a broader variety of disease research projects including those with significant impacts to the UK and some developing countries, such as mental disorders and neglected tropical diseases. The NSFC tends to fund projects that focus on the diseases with the highest burden to China, such as different kinds of neoplasms, but not to rapidly-spreading diseases, such as HIV/AIDS and sexually transmitted infections, and mental disorders.

Introduction

The Global Observatory on Health R&D of the World Health Organization, a recent initiative that aims to help identify health R&D priorities based on public health needs, tries to bring together health information and statistics and research information and statistics. The initiative is in need of advanced informetric methods. At the same time, it gives informetrics an opportunity to contribute to societal development. Among the latest top 100 most-mentioned scholarly articles on Altmetrics.com of 2018, 44 articles are related to medical and health science. Of these, one third are concerned with the global burden of disease (Griswold et al., 2018). Both researchers and the general public have shown an increasing interest in the field of health, given rising concerns over health issues and the growing need to allocate health research investments in line with public demand (Atala, Trinquart, Ravaud, & Porcher, 2018; Røttingen et al., 2013). Early in 2012, the WHO Consultative Expert Working Group published a report that focuses on publicly-funded research into the diseases carrying the highest burden for developing countries (World Health Organization (WHO), 2012b). China, as the largest developing country and the world's second-biggest economy with rapid economic and technological development, is in a unique position that deserves a detailed investigation into the issue of whether the diseases with the highest burden have received sufficient attention from government-funded research.

As an essential channel for supporting independent research, scientific funding is highly prized in China. Thereinto the National Natural Science Foundation of China (NSFC) is the leading and the largest funding agency for basic research in China (Wu, Yuan, Li, & Li, 2018). Previous research indicates that China has the leading average funding ratio in the world: 70.34% of SCI/SSCI indexed papers are supported by some level of research funding, and almost 90% of those are supported by the NSFC (Wang, Liu, Ding, & Wang, 2011). The NSFC has made great strides in promoting basic research into natural science, especially medical and health-related research. For this reason, we selected the NSFC as the main research object of this study. To conduct a comparative analysis, we selected the Medical Research Council of the UK (MRC) as a second research object. The MRC is a crucial sub-agency within the UK Research and Innovation (UKRI) with a particular focus on coordinating and funding research into medical and health science.

With such large investments being made into health research every year, assessing the efficiency and sufficiency of funded research is a valuable undertaking for governors, funding organizations, and academia – and not simply in terms of the number of publications and citations but also from the perspective of the topics and content being funded. A combined perspective on which diseases should be receiving research funding alongside those that are can give decision makers clear recognition of their current funding strategy. Such insights may suggest continued support of current endeavors, modifications to existing policies, or demand new strategies to improve scientific development and make research activity more relevant and more effective (Ebadi & Schiffauerova, 2016).

In fact, there are some studies have been conducted to provide in-depth analysis of disease and its effects on society. For example, Begum et al. (2018) mapped research activity into cancer in 29 countries during the period of 2007-2016. Kalita, Shinde, and Patel (2015) used bibliometric analysis to describe the focus and distribution of public health research output in India, finding marked inequities in relation to the burden of disease and the geographic distribution of research. While consummate, the research to health science and funding policies lacks a combination of these two perspectives – i.e., what is being funded and what should be funded – to explore whether or not funded scientific research meets social demands. Therefore, beyond identifying how research into the hottest currently-funded topics has evolved, a comparative analysis between research focus and high-burden diseases is also needed to explore the relationship between health research investment and public health demand. That is the focus of this paper.

Data

Funding Data

Most previous bibliometric studies derive funding information from publications. But such an approach may cause problems with incomplete and inaccurate data that may skew the study's results (Tang, Hu, & Liu, 2017). In this study, we acquired funding data directly from official funding organizations to ensure better accuracy and reliability of the data.

The NSFC funding system provides three categories of programs: Research Programs, Talent Training Programs, and Research Support Programs, which sponsor cooperative international research. The various research disciplines are managed between eight scientific departments. Falling within the Research Programs category, we chose to study "General Programs" because

these are the most fundamental projects and receive the lion's share of total funding, and projects funded by the Department of Health Sciences were chosen as the main research scope as this is the department responsible for managing health and medical research funding. (Note that, prior to 2010, such projects were managed by the Department of Life Sciences, but the project codes between the two departments are consistent across the whole period of study. These codes were used as the basis for selecting projects to include in the corpus.)

The MRC is one of seven councils that comprise UK Research and Innovation (UKRI) and is responsible for co-coordinating and funding medical and health research. UKRI was established in 2018 as a new umbrella body to replace Research Councils UK (RCUK), but both devote to coordinate funding and academic research for research councils. Previously, in addition to seven research councils across the arts, humanities, sciences and engineering, two new councils have been added to UKRI, which are Research England and Innovate UK. Each Research Council is an individual non-departmental government agency that receives funding from the UK government's science budget to fund research in different areas. Like the NSFC's General Program, MRC mainly funds research grants, which were selected as the research objects for this study.

Python web crawler technology was used to collect raw data from the NSFC official websites (http://npd.nsfc.gov.cn/;https://isisn.nsfc.gov.cn/egrantindex/funcindex/prjsearch-list) and the Gate way to Research of UKRI (https://gtr.ukri.org/search/project?term=*). The last time the NSFC website was updated in December 2018 and the Gate way to Research of UKRI updated irregularly; we acquired the data on 26 December 2018. The dataset comprised projects titles, abstracts, principal investigators, granted organizations, and granted dates from the NSFC's Department of Health Sciences General Program and the MRC Research Grants program for the period 2006 to 2017. The dataset was further divided into three time periods of four years each for the dynamic analysis. The initial dataset contained 38,216 NSFC-funded and 4202 MRC-funded projects, of which 1387 and 195 projects were removed due to lack of an abstract or title. Thus, the final dataset contained a total of 36,829 NSFC-funded projects and 4007 MRC-funded projects. It is worthy of note that no abstracts were available for the NSFC-funded projects for the period 2014 to 2017. Therefore, we used the project titles as substitutes for the abstracts for this time interval. The annual number of funded projects for both funding organizations during 2006 -2017 are shown in Figure 1.

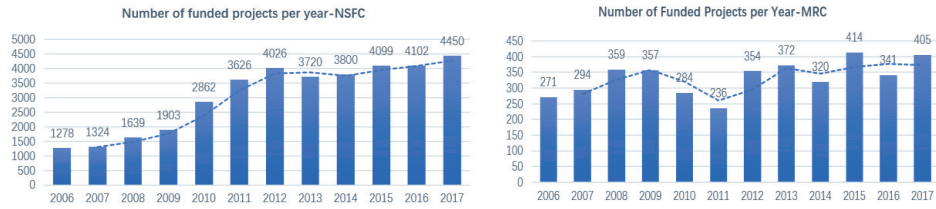


Figure 1. The annual trends of projects funded by the NSFC and the MRCs during 2006-2017

Burden of Disease

The burden of disease data used in the comparative analysis was sourced from the Global Burden of Disease (GBD) (http://www.healthdata.org/gbd/about), which is a comprehensive regional and global research program of disease burden that assesses mortality and disability from major diseases, injuries, and risk factors. GBD is based out of the Institute for Health Metrics and Evaluation (IHME)

at the University of Washington, and it also was institutionalized at the World Health Organization (WHO). With the aim of measuring disability and death from a multitude of causes worldwide, the GBD study attributes each death to a single underlying cause that began the series of events leading to death, in accordance with the International Classification of Diseases (ICD) (GBD 2017 Causes of Death Collaborators, 2018). Hence, our comparative analysis follows these principles. The hierarchy consists of four levels. At the highest level (Level 1), all disease burden is divided among three mutually exclusive and collectively exhaustive categories: communicable, maternal, neonatal, and nutritional; noncommunicable diseases (NCDs); and injuries. Level 2 distinguishes these Level 1 categories into several cause groups, and Levels 3 and 4 further disaggregate these causes. We mostly focused on the Level 2 and 3 categories for our analysis.

Analysis Method and Tools

Several important procedures were used to conduct this research, which are shown in Figure 2.

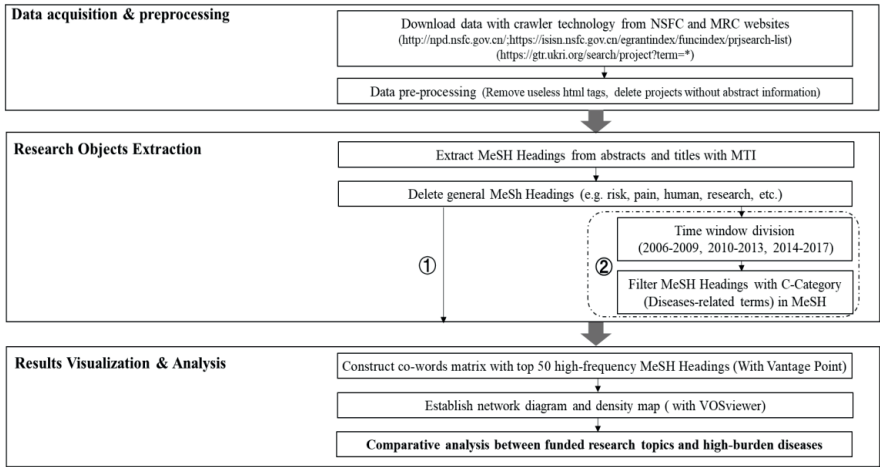


Figure 2. Research flow diagram for funded projects topic analysis

Notes: ① indicates the general analysis research procedure (Section 4), which outlines the topic skeleton of the research.

② indicates the procedure used to undertake the comparative analysis over different time intervals in the same section.

Term interpretation: **MeSH** stands for Medical Subject Headings, which is a biomedical indexing vocabulary maintained by the U.S. National Library of Medicine (NLM); **MTI** stands for Medical Text Indexer (MTI), which provides indexing recommendations based on MeSH; **MeSH Headings** means the recommended headings produced by MTI from the projects' abstracts and titles.

We used the NLM Medical Text Indexer (MTI) (https://ii.nlm.nih.gov/MTI/index.shtml) to extract the corresponding MeSH Headings from the abstracts and titles of the projects. MeSH is a biomedical indexing vocabulary maintained by the US NLM, and MTI is the main product of the NLM's Indexing Initiative project. MTI is able to summarize input text into an ordered list of MeSH Headings, which we used as the unit of analysis for this study. Further, we chose to use the MeSH Headings as recommended by MTI rather than the keywords provided by the applicant for a number of reasons. First, because project keywords are provided by applicants, they have relatively higher subjectivity. Plus, they may not always accurately reflect the research area, especially for specific diseases. Second, keywords are not standardized, and it is common for applicants to use different expressions for the same disease. In co-word analysis, the bias this creates is problematic. MTI

offers a basic standard for extracting MeSH Headings from abstracts and titles, which forms a more precise and accurate basis of analysis.

To construct the topic matrix of funded projects over the whole period (2006-2017) for the general analysis, we calculated the frequency of each topic using the MeSH Headings as is from Step 2. However, for the comparative analysis across different time intervals, we filtered the MeSH Headings by their C-category. The C-category segregates diseases, so only funded projects relating to disease were analyzed in further detail. The top 50 MeSH Headings were considered as high-frequency terms. The term frequency was calculated as the co-occurrence of two terms appearing in the same project to produce a symmetrical co-word matrix, which was later used in the clustering and visualization analysis generated by VOSviewer.

Results

As previously mentioned, our results derive from two separate analyses: 1) general analysis: outlines the topic structure of funded research over the whole period (2006-2017); 2) comparative analysis: provides a "heat map" of research into clusters of various diseases in a series of time intervals in comparison to the burden of those diseases.

General Analysis

Figure 3 shows the top 50 most frequently funded MeSH Headings between 2006 and 2017 by the NSFC according to relative size of the main health science research topics. Signal transduction has been the most-funded topic for the NSFC for the last 12 years, which falls within the cytokines domain and is relevant to a variety of diseases, such as cardiovascular diseases (Atherosclerosis) and diabetes (Oxidative Stress) as shown in the cluster at top left part of Figure 3. This network map makes the links between abnormal signal transduction and NCDs clear. One surprising point in this cluster is "Medicine, Chinese Traditional". From this, we surmise that the NSFC places some importance on funding the characteristic Chinese medical system and this is likely due to a desire by the State Council to integrate traditional Chinese medicine with Western medical protocols.

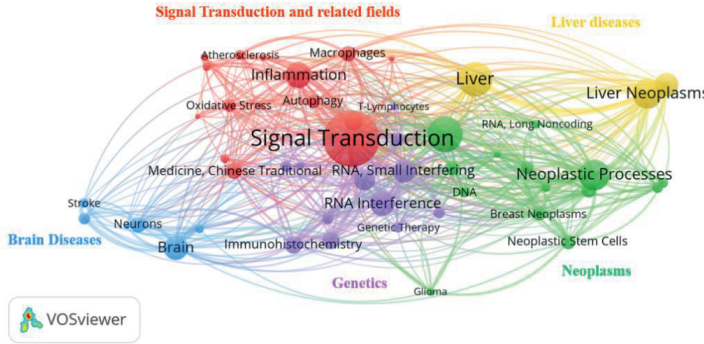


Figure 3. The occurrence map of MeSH Headings in NSFC funded research during 2006-2017

Another obvious cluster points to neoplasm-associated disease. Liver and liver neoplasms is an adjunct cluster, illustrating that liver diseases are a consistent concern for researchers of funded projects. Genetics has also received great attention from NSFC; and our dataset demonstrated the association of genetics with various diseases. However, it is hard to investigate the concrete connection between specific disease with genetics research due to the limitation of graphic

visualization (Figure 3), which requires further analysis. Brain-related diseases and neurological disorders form another cluster of NSFC-funded research.

Figure 4 illustrates the topic structure of the MRC-funded projects. NCDs, such as neoplasms, diabetes, and several other MeSH Headings related to cardiovascular diseases form one cluster. But, unlike the prime focus of NSFC funding, the most profound cluster concerns infectious diseases. Within infectious diseases, HIV and its related illnesses have attracted the most attention. Another obvious difference is that the MRC has paid more attention to mental diseases than the NSFC. In fact, depression and schizophrenia do not even feature on the NSFC map.

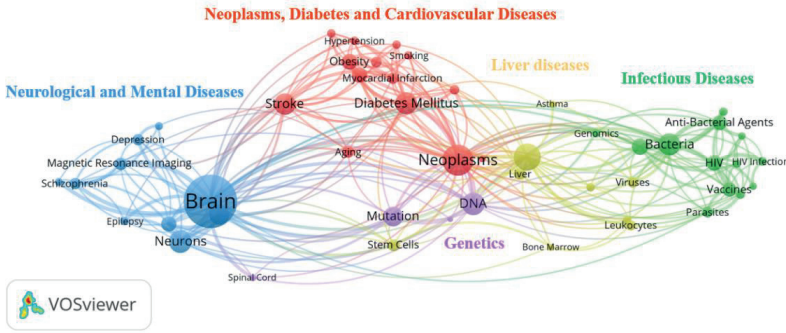


Figure 4. The occurrence map of MeSH Headings in MRC funded research during 2006-2017

Another important term to note is "aging". In 2006, more than 16% of the UK's population was over the age of 65, while the universally-recognized standard for an aging society begins at around 7% (The World Bank, 2017). Compared to China with an aged population of 7.84%, the UK is clearly grappling with a severe aging problem. Interestingly, the occurrence of neoplasms and other diseases increases with age (Torre et al., 2015), which would explain the close attention MRC has paid to these types of diseases as illustrated by the cluster in the middle of Figure 4.

Comparative Analysis with Disease Burden

Analysis of the Top 5 Disease with the Highest Burden

The GBD is a concept the WHO has been promoting for over a decade under its express mandate to report on health information. The GBD approach results in a summary measure of morbidity, disability, and mortality as one single number – disability-adjusted life years (DALYs) (GBD 2017 Causes of Death Collaborators, 2018). DALYs indicates the time lost through premature death and the time lived in a state of less than optimal health, loosely referred to as "disability". One DALY can be thought of as one lost year of "healthy" life. For a specific cause of death or disability, DALYs are calculated as the sum of the years of life lost (YLLs) plus the years of reduced quality of life (disability) (YLDs) as a direct result of that cause, i.e., DALY=YLL+YLD. Following the WHO, we used this DALY formula in the following comparative analysis as the index for the burden of disease.

Table 1 lists the top five diseases classified at Level 2 of GBD cause of death list. Four of these five diseases have the highest burden index assessed by DALYs per 100,000 across all three time periods under study in both China and the UK. Hence, the two countries share a similar structure for the burden of disease. Chronic respiratory disease in China and neurological disorders in the UK are the exceptions. The top two diseases for both countries are cardiovascular disease and

neoplasms. Cardiovascular disease has the highest DALYs per 100,000 in China and neoplasms rank highest in the UK. In addition, the ranks for these five diseases have, for the most part, remained relatively stable. In China, the incidence of musculoskeletal disorders has increased across the full period. One possible reason may be that China's change in economy, and therefore living and working habits, has increased the number of work hours in sedentary occupations. Indeed, the main contributors to the rise in musculoskeletal disorders are neck pain and lower back pain, which are classified at Level 3 of the ICD, and increase rates of DALYs per 100,000 from 2006 to 2017 is 12.26% for neck pain and 9.13% for low back pain.

Table 1. Top 5 Diseases with the highest DALYs per 100,000 in China and the UK

Region	China		
Period	**2006-2009**	**2010-2013**	**2014-2017**
Top 1	Cardiovascular diseases	Cardiovascular diseases	Cardiovascular diseases
Top 2	Neoplasms	Neoplasms	Neoplasms
Top 3	Chronic respiratory diseases	Chronic respiratory diseases	***Musculoskeletal disorders***
Top 4	Musculoskeletal disorders	***Musculoskeletal disorders***	Chronic respiratory diseases
Top 5	Mental disorders	Mental disorders	Mental disorders
Region	**the UK**		
Period	**2006-2009**	**2010-2013**	**2014-2017**
Top 1	Neoplasms	Neoplasms	Neoplasms
Top 2	Cardiovascular diseases	Cardiovascular diseases	Cardiovascular diseases
Top 3	Musculoskeletal disorders	Musculoskeletal disorders	Musculoskeletal disorders
Top 4	Neurological disorders	Neurological disorders	Neurological disorders
Top 5	Mental disorders	Mental disorders	Mental disorders

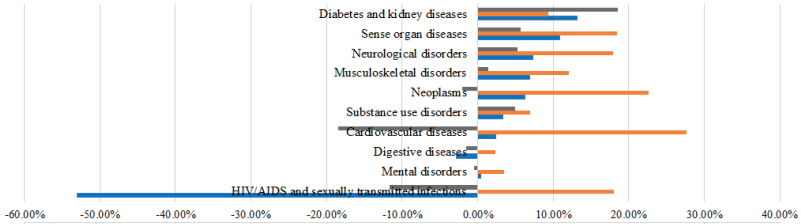
Note: Data from Global of Burden of Disease https://vizhub.healthdata.org/gbd-compare/

Another notable point is that mental disorders ranked in the top five for both countries. According to Level 4 of the GBD cause of death list, major depressive disorder is the leading contributor to this high rank (In China, the DALYs increased by 6.19% across the period, and 0.57% for the UK.) However, according to the results of our general analysis, only the MRC has devoted significant funding to research projects concerning mental disorders. Unfortunately, no terms relating to mental disorders exist in the C-category of MeSH, except for "mental fatigue", which precludes exploring this observation further given the data used for this comparative analysis.

Analysis of the Change in Burden of Disease Rates

Taken together, all ten of the top diseases with the highest DALYs per 100,000 have increased across China, the UK, and the rest of the world over the last 12 years. We selected ten diseases with significant rates of change for further analysis. Table 2 charts the rate of change in disease burden for various regions.

Table 2. Annual percentage change of DALYs per 100,000 from 2006 to 2017, both sex & all ages



	HIV/AIDS and sexually transmitted infections	Mental disorders	Digestive diseases	Cardiovascular diseases	Substance use disorders	Neoplasms	Musculoskeletal disorders	Neurological disorders	Sense organ diseases	Diabetes and kidney diseases
United Kingdom	-11.64%	-0.42%	-1.48%	-18.45%	4.94%	-2.03%	1.43%	5.29%	5.67%	18.61%
China	18.07%	3.47%	2.40%	27.63%	6.92%	22.67%	12.12%	17.98%	18.47%	9.40%
Global	-53.02%	0.46%	-2.81%	2.42%	3.32%	6.28%	6.87%	7.31%	10.92%	13.22%

■United Kingdom ■China ■Global

Note: Burden data from Global of Burden of Disease https://vizhub.healthdata.org/gbd-compare/

The density maps of diseases with high attention from funded research for all three time-intervals, provided in the Appendix, reveal that three major NCDs – neoplasms, diabetes, and cardiovascular disease – are the most common research concerns for both the NSFC and the MRC. This concern is consistent with the level of disease burden for both countries (see Table 1). The Diabetes and kidney diseases category has also significantly contributed to increases in the mortality rate over the past decade (GBD 2017 Causes of Death Collaborators, 2018), with a strong upward trend in all regions as shown in Table 2.

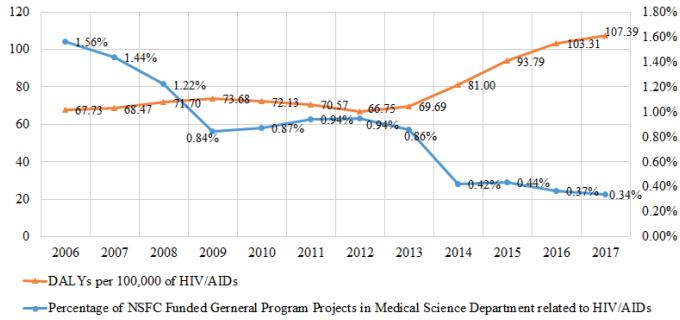


Figure 5. Percentage of funded projects relating to HIV/AIDs infections and the burden of HIV/AIDs infections

Note: Burden data from Global of Burden of Disease https://vizhub.healthdata.org/gbd-compare/

Yet, beyond these shared concerns, the research into more specific diseases funded by the NSFC and the MRC exhibit substantially different focuses. In China, the NSFC has funded more research into lung, liver, testicular, and colonic neoplasms. Additionally, Alzheimer's disease has remained a consistent subject of funding for the NSFC as the main contributor to the increasing rate of neurological disorders. Digestive diseases, HIV/AIDs, and sexually transmitted diseases (STD) show an increasing burden in China but tend to be decreasing in the UK and the rest of the world. While the NSFC does appear to be funding research into some digestive diseases, particularly liver cirrhosis, it is striking that the MeSH Headings relating to HIV/AIDs and STD do not appear in China's high-frequency list or the subsequent matrix.

Multiple factors may exert an influence over the change in burden for HIV/AIDs and STDs, it is difficult to argue any causality between the burden and the number of relevant funded projects. Moreover, no specific funded category in the NSFC Department of Health Sciences can be directly matched with HIV/AIDs or STDs, except for one small category called 'sexually transmitted infections' with code H1910 (refer to the Application codes of NSFC, http://www.nsfc.gov.cn/publish/portal0/tab558/). But only 0.09% of NSFC-funded projects were found in this category (32 of 36,829), which must reflect the lack of consideration being paid to STDs by the NSFC and possibly even medical academia, at least in part. Another possible explanation is that some research projects are relevant to HIV/AIDs and STDs but are not included under H1910. To further investigate this possibility, we conducted a detailed search of the funded projects relating to HIV/AIDs and STDs, the number of funded projects related to HIV/AIDs infections has been calculated in terms of counting the number of projects containing "HIV or AIDs" in its recommended MeSH Headings by MTI (as explained in *Term interpretation*

of Figure 2), and we use the percentage of funded projects related to HIV/AIDs infections and the burden of HIV/AIDs infections to make comparison analysis. According to Figure 5, the percentage of projects funded by the NSFC related to HIV/AIDs and STDs was not only relatively low but has seen a downtrend with obvious drops between 2008 and 2009 and in 2013 to 2014. Notably, there was a slight increase in funding from 2009 to 2012, but the DALYs per 100,000 for these diseases decreased during this period. Compared to the decreasing burden of these diseases across the world and in the UK, China has seen a sharp increase in DALYs per 100,000 caused by HIV/AIDs since 2013. In China, the three main subpopulations affected are drug users (Wang et al., 2015), female sex workers (Wang et al., 2014) and men who have sex with men (Chow et al., 2014). More than 90% of new HIV/AIDs infections incidences were transmitted through sex from January to October in 2014, according to a news report by Chinanews.com and data from the National Center for AIDS/STD Control and Prevention of China (China News Service (CNS), 2014). The risk of contracting HIV/AIDs through same-sex relations between men in China is exceptionally high with a prevalence rate that rose from 5.73% in 2010 to 7.75% in 2014 (Cui et al., 2016). These statistics should be sounding alarms to both the National Government and Chinese society, but it seems they are not.

In addition to HIV/AIDs and sexually transmitted infections, the NSFC has paid scant attention to mental disorders based on the above analysis. From the perspective of the evolution of which research topics receive funding (see Appendices), there were only slight changes between the three time-intervals.

MRC-funded research has also placed a large spotlight on neoplasms, and attention on neurological disorders, especially Alzheimer's disease; Parkinson's disease; and epilepsy, has gradually increased. Moreover, the MRC appears to fund a larger variety of disease research beyond these three main focuses compared to the NSFC, such as blindness, hearing loss, and other diseases of the sensory organs, and musculoskeletal forms of arthritis. The MRC has also placed heightened attention on HIV/AIDs, malaria, parasitic diseases, and other types of infections.

The Commission on Macroeconomics and Health divides diseases into three types according to the national income level and the burden of disease (World Health Organization (WHO), 2012a). Type I includes diseases of incidence in both developed and developing countries. Type II also includes diseases that affect both types of country, but developing countries substantial more so. Type III refers to diseases that are overwhelmingly or exclusively present in developing countries. For example, tuberculosis and diarrhea are considered to be Type II diseases, while malaria is a Type III disease. Our findings show that the MRC has placed more emphasis on funding disease research in the Type II and III categories than the NSFC. Correspondingly, the overall burden of communicable causes of death have decreased over the last decades, and the largest contributors to this decrease include reduced DALY rates for HIV/AIDS, tuberculosis, diarrhea, and malaria (GBD 2017 Causes of Death Collaborators, 2018)– all of which have been consistent topics of funding for the MRC. In terms of the evolution of funding focus (see Appendix), the MRC has broadened its scope of project funding to suit research "hotspots" at different times.

Conclusion and Discussion

Summary

In this work, we explored the focus of disease-related research projects by national funding organizations and juxtaposed levels of funding with the burden of disease. Our analysis compares China and the UK using co-word and network analysis on funded project data from China's NSFC and the UK's MRC.

We find that both the NSFC and the MRC have devoted significant funding to the top two diseases with the highest national burden – neoplasms and cardiovascular disease. Diabetes and kidney disease have also received substantial attention, which corresponds to an increasing trend in burden in both countries from 2006 to 2017. One of the major differences between the types of projects sponsored by the NSFC and the MRC is the variety of diseases. MRC-funded research topics span more kinds of diseases, including some neglected tropical diseases, while the NSFC has mainly focused on the three diseases mentioned above. One explanation for this could be the different funding requirements of each organization. Only Chinese scholars can apply for the NSFC General Program, while scholars from all over the world can be awarded an MRC research grant.

The UK's burden of disease is relatively stable compared with China, which may due to a more developed and stabilized social development structure. The variety of burdensome diseases is growing in China, but are not all are receiving adequate attention from the national funding organization, especially HIV/AIDs and STDs. Notably, China is undergoing a period of social transformation, where competition, the speed of one's life rhythm, and heavy working pressure are not yet fortified in the social conscience. This may be one reason for the increasing incidence of various diseases and the aggravated burden of disease. Paradoxically, research into mental disorders is easily overlooked by individuals and government and requires more consideration from both a social and a scientific perspective.

Limitation and Future Research

This study contains two data-related limitations that need to be addressed. The first one is our comparative analysis only included data from the division of scientific departments within the NSFC and the division of Research Councils in the UK, where medical and health research and bio-scientific research are divided and managed separately. We drew our data corpus from the Department of Health Sciences (NSFC, China) and the MRC, UK, which both manage medical and health-related research. However, it is inevitable that some relevant research projects would have been assigned to the agencies responsible for bio-scientific research, i.e., the BBSRC in the UK and the Department of Life Sciences in China.

The second one is that we did not link the MeSH Headings with the principles of the GBD cause of death list and ICD, because there is no standardized classification system to do this. Therefore, the above results cannot be analyzed in the context of a unified classification. Further collaborative research between academics in medical science and informetrics would be required to build such a system.

Acknowledgements

This work is supported by the National Natural Science Foundation of China: Grants 71573085, the Excellence Scholarship in Social Science in HeNan Province (No.2018-YXXZ-10) and National Laboratory Center for Library and Information Science in Wuhan University.

References

Atala, I., Trinquart, L., Ravaud, P., & Porcher, R. e. (2018). A mapping of 115,000 randomized trials revealed a mismatch between research effort and health needs in non-high. *Journal of Clinical Epidemiology, 98*, 123-132.

Begum, M., Lewison, G., Jassem, J., Mixich, V., Cufer, T., Nurgozhin, T., et al. (2018). Mapping cancer research across Central and Eastern Europe, the Russian Federation and Central Asia: Implications for future national cancer control planning. *European Journal of Cancer, 104*, 127-136.

China News Service (CNS). (2014). 87,000 new HIV/AIDs cases were report from January to October in 2014 in China, more than 90% of new HIV/AIDs infections incidences were transmitted through sex (In Chinese). Retrieved December 02, 2018, http://www.chinanews.com/gn/2014/12-01/6832929.shtml

Chow, E. P., Lau, J. T., Zhuang, X., Zhang, X., Wang, Y., & Zhang, L. (2014). HIV prevalence trends, risky behaviours, and governmental and community responses to the epidemic among men who have sex with men in China. *BioMed Research International, 2014*, 607261.

Cui, Y., Guo, W., Li, D., Wang, L., Shi, C. X., Brookmeyer, R., et al. (2016). Estimating HIV incidence among key affected populations in China from serial cross-sectional surveys in 2010-2014. *Journal of the International AIDS Society, 19*(1), 2010-2017.

Ebadi, A., & Schiffauerova, A. (2016). How to boost scientific production? A statistical analysis of research funding and other influencing factors. *Scientometrics, 106*(3), 1093-1116.

GBD 2017 Causes of Death Collaborators. (2018). Global, regional, and national age-sex-specific mortality for 282 causes of death in 195 countries and territories, 1980-2017: a systematic analysis for the Global Burden of Disease Study 2017. *Lancet, 392*(10159), 1736-1788.

Griswold, M. G., Fullman, N., Hawley, C., Arian, N., Zimsen, S. R. M., Tymeson, H. D., et al. (2018). Alcohol use and burden for 195 countries and territories, 1990–2016: a systematic analysis for the Global Burden of Disease Study 2016. *Lancet, 392*(10152), 1015-1035.

Kalita, A., Shinde, S., & Patel, V. (2015). Public health research in India in the new millennium: a bibliometric analysis. *Global Health Action, 8*(1), 27576.

Røttingen, J.-A., Regmi, S., Eide, M., Young, A. J., Viergever, R. F., Årdal, C., et al. (2013). Mapping of available health research and development data: what's there, what's missing, and what role is there for a global observatory? *Lancet, 382*(9900), 1286-1307.

Tang, L., Hu, G., & Liu, W. (2017). Funding acknowledgment analysis: Queries and caveats. *Journal of the Association for Information Science and Technology, 68*(3), 790-794.

The World Bank. (2017). Population ages 65 and above (% of total population). Retrieved November 16, 2018, https://data.worldbank.org/indicator/SP.POP.65UP.TO.ZS?end=2017&locations=CNGB&start=1960&view=chart

Torre, L. A., Bray, F., Siegel, R. L., Ferlay, J., Lortet-Tieulent, J., & Jemal, A. (2015). Global cancer statistics, 2012. *CA: A Cancer Journal for Clinicians, 65*(2), 87-108.

Wang, L., Guo, W., Li, D., Ding, Z., McGoogan, J. M., Wang, N., et al. (2015). HIV epidemic among drug users in China: 1995-2011. *Addiction, 110* (Suppl. 1), 20-28.

Wang, L., Tang, W., Wang, L., Qian, S., Li, Y. G., Xing, J., et al. (2014). The HIV, syphilis, and HCV epidemics among female sex workers in china: results from a serial cross-sectional study between 2008 and 2012. *Clinical Infectious Diseases, 59*(1), e1-9.

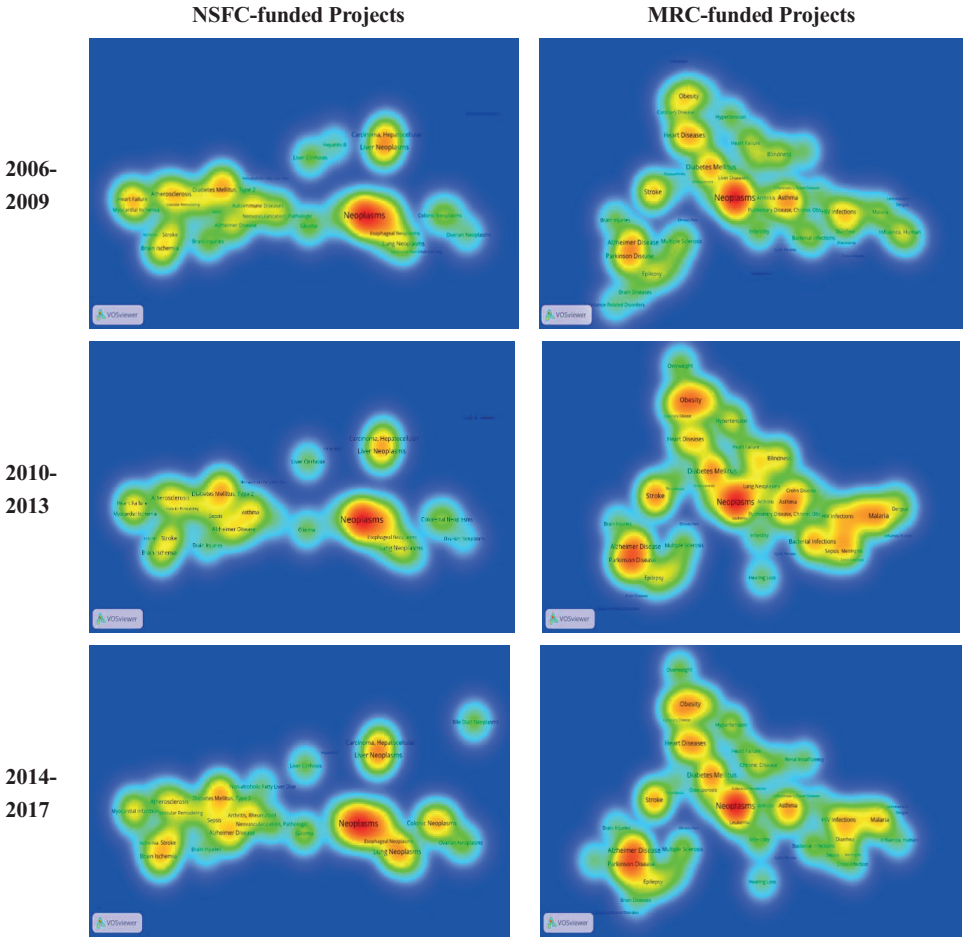
Wang, X., Liu, D., Ding, K., & Wang, X. (2011). Science funding and research output: a study on 10 countries. *Scientometrics, 91*(2), 591-599.

World Health Organization (WHO). (2012a). *Background document provided by the WHO Secretariat-Defining disease types I, II and III*. Retrieved from https://www.who.int/phi/3-background_cewg_agenda_item5_disease_types_final.pdf

World Health Organization (WHO). (2012b). *Report of the Consultative Expert Working Group on Research and Development: Financing and Coordination*. Retrieved from https://www.who.int/phi/cewg_report/en/

Wu, D., Yuan, L., Li, R., & Li, J. (2018). Decomposing inequality in research funding by university-institute sub-group: A three-stage nested Theil index. *Journal of Informetrics, 12*(4), 1312-1326.

Appendix



Appendix 1. Density map of diseases with high attention: Co-words clustering of funded projects' abstracts from NSFC and MRC

Notes: the color scale represents the relative frequency of papers containing a specific term, from low-frequency (blue) to high-frequency (red). The clustering was computed using the VOSviewer algorithm based on a co-word matrix that was constructed using the procedure described in the "Analysis methods and tools" section. Each node represents an obtained MeSH Heading.

Discovering types of research performance of scientists with significant contributions

Yu-Wei Chang[1] and Mu-Hsuan Huang[2]

[1] *yuweichang2013@ntu.edu.tw*
National Taiwan University, Dept of Library and Information Science, Taipei 10607 (Taiwan)
National Taiwan University, Center for Research in Econometric Theory and Applications, Taipei 10617(Taiwan)

[2] *mhhuang@ntu.edu.tw*
National Taiwan University, Dept of Library and Information Science, Taipei 10607 (Taiwan)

Abstract

This study compared the longitudinal research performance of 50 biological scientists who have received the National Medal of Science (NMS) between 2005 and 2014. The results show that each scientist had shared the honor of receiving the fellowship of American Academy of Arts and Sciences (AAAS). The average age of those obtaining the AAAS fellowship was 50 years, whereas that of those receiving NMS was 69 years. The 50 scientists were categorized based on seven types of research performance in terms of research productivity and research influence. Based on the annual number of publications over the course of three periods, 10 types of research productivity were identified. The primary type of research productivity indicated an upward trend in the first and second periods, whereas the third period indicated a decreasing trend. Based on the annual average citations received per publication, 11 types of research influence were distinguished. The primary type of research influence indicated an increasing trend in the first period, whereas the second and third periods indicated a decreasing trend. Among the participants, 19 scientists had similar trends in both research productivity and research influence. Few scientists' significant scientific contributions were presented in influential books and were used to assist humans. Highly-cited publications were noted as being effective for identifying excellent scientists.

Introduction

Numerous prestigious scientific awards aim to acknowledge scientists who have significantly contributed to the literature. The winners of these awards can be defined as excellent scientists. Although the factors that define a significant scientific contribution have not been outlined (Khosravi & Chavan, 2012), the scope of these contributions exceeds the number of publications and those of citations received through publications, according to assessment criteria that were announced by distinguished scientific awards. Indicators related to the numbers of publications (research productivity) and those of citations received through publications (research influence) are widely used to measure the research performance of individual researchers (Abramo, D'Angelo, & Solazzi, 2011). Scientists' publications mark their scientific contributions and represent the results of their research. Identifying outstanding researchers is one purpose of research performance, which is same as that of establishing a scientific award. Therefore, the association between scientific contributions and research performance is a notable factor for analysis.

Studies have analyzed the differences in characteristics in research productivity and research influence between recipient and nonrecipient scientists of a specific award (Borjas & Doran, 2015; Chan, Frey, Gallus, & Torgler, 2014) or focused on only the research productivity of winners of scientific awards (Licea De Arenas, Valles, & Arenas, 1999; Yair, Gueta, & Davidovitch, 2017). Chan et al. (2014) observed the increase in research productivity and research influence among award winners when they obtained awards early in their academic

career. However, Borjas and Doran (2013) revealed the decrease in publication rates among the winners of the Fields Medal. The Fields Medal is the most prestigious award in the field of mathematics and is awarded to mathematicians under 40 years of age. Therefore, the age of the scientists receiving excellence awards and the status of the awards will affect their subsequent publication productivity. Moreover, some studies measured only the award winners' research productivity or both their research productivity and the number of citations received within a specific period (Licea et al., 1999; O'Connell & Rugman, 2013). The research performance of individual scientists with notable scientific contributions over the course of their academic career have not been examined in the extant literature.

Scientists with excellent achievements are believed to have increased research productivity and research influence compared with their peers; therefore, they are also believed to share similar characteristics in research performances. However, further research is required to determine whether excellent scientists have noteworthy research performance. This study primarily examines whether scientists with excellent scientific contributions have similar research performance in terms of research productivity and research influence at the individual level. Although award winners were not selected on the basis of only research productivity and research influence, this study focused on determining whether the scientific contributions of these scientists can be observed using their research performance. If similar research performance is noted among individual excellent scientists, this may imply that research performance is an essential factor for indicating their scientific contributions. However, if dissimilar research performance is noted among these scientists, this may imply that the scientific contributions that are noticed by award reviewers cannot be reflected substantially through research performance. This may also imply that excellent scientists have various types of research performance with diverse characteristics.

The longitudinal research performances of scientists with significant scientific contributions can help to explore the differences between research performance and scientific contributions. Changes in the annual number of publications and the annual average citations received for each publication determined the types of research productivity and research influence for each scientist. The differences between the types of research productivity and that of research influence further formed the types of research performances. This study addresses the following research questions:

1. Do excellent scientists have similar types of research productivity?
2. Do excellent scientists have similar types of research influence?

Methodology

Data collection

We defined excellent scientists with scientific contributions as receiving at least three international and national scientific awards and honors. This means that numerous professional organizations and experts have recognized their scientific contributions. Considering the level, reputation, and history, we selected the winners of National Medal Science (NMS). NMS was established by the US Congress in 1959 and is currently administrated by the National Science Foundation. Since 1962, the medal's winners have been selected based on their field (behavioral and social sciences, biology, chemistry, engineering, mathematics, and physics) and have been awarded their medals by the US President. NMS candidates must first be nominated, following which they must submit less than 10 publications and patents with individual contribution statements. A criteria for selecting winners includes the influence of the publication and its contributions to the scientific field, society, education, industry, and the country (National Science Foundation, 2018). Considering the disciplinary difference and the total number of winners, 50 biological scientists who had received the NMS during a recent 20-year period

(1995–2014) were selected as the subject of this study. No NMS winners are announced after 2015.

Data was collected by using the winner's names, affiliations, and their brief statement about winning the medal, background information regarding their research output, awards and honors, education, and job experience were collected through their curricula vitae, personal websites, and Internet resources. Because the complete publication lists for each scientist cannot be obtained, the scope of publications was limited to publications indexed by two large interdisciplinary databases of Web of Science (WoS) and the Scopus. Considering the differences in the characteristics of the various types of publications, only articles, conference papers, and review articles were deemed as the primary types of research publications, thus becoming the sample publications used in this study. The other type of documents that were most commonly found were editorial materials. No books were retrieved through WoS and only a few books were retrieved from the Scopus for some of the scientists. The first year for counting publications from each scientist varied based on their first publication, whereas the final year was 2017. The research influence of publications referred to the number of citations received by articles, review articles, and conference papers from WoS, instead of the Scopus, because the annual number of citations that were added to individual publications were available from WoS for further analysis. The final year for counting the annual number of citations received by publications was also 2017. The bibliographical records of publications obtained from the two databases were used for related analyses. Each bibliographical record included the title, author name and affiliation, document type, publication year, source, and total number of citations. The annual number of citations that were received by the sample publications were available and collected in a separate file.

Data processing

Two indicators were used to measure the research performance of each scientist. The first indicator, annual research productivity, was used to measure the annual number of publications for each scientist. The annual number of publications covered by WoS, those outside of WoS, and those covered by Scopus were counted separately. Therefore, the bibliographic records of publications obtained from WoS and Scopus were compared, and publications obtained outside the WoS were identified. The second indicator, annual research influence, referred to the annual average citations received for each WoS publication.

After reviewing the background information of 50 biological scientists, each scientist was found to have obtained several awards and honors including the fellowship granted by the American Academy of Arts and Sciences (AACS). Scientists who receive the AACS fellowship are widely considered to be outstanding and influential. Moreover, most scientists have received the NMS after they were past the age of 65 years. This implies that bulk of their research performance in their academic careers occurred before they received the NMS. Therefore, we decided to elaborate the focus of their research by observing changes in their research performance during three periods, namely before obtaining the AACS fellowship, after obtaining the AACS fellowship and before receiving the NMS, and after receiving NMS. The changes in research productivity per year and those in average research influence per article per year during each period deduced the specific types of research productivity and that of research influence. Code 0 referred to no trends revealed; code 1 referred to an increasing trend; and code 2 referred to a decreasing trend. The samples indicating no trends include figures related to research productivity and research influence that include less than 3 years in a single period. Therefore, each type was labeled with three numbers. For example, type 111 referred to an increasing trend in three periods, whereas type 122 implied that an increasing trend was noted in the first period and a decreasing trend was noted in the second and third periods. In the study sample, 10 types of research productivity and 11 types of research influence were observed.

The differences between the types of research productivity and that of research influence for each scientist were used to form seven broad types of research performance.

Results

Annual average research performance for each scientist

Table 1 shows that the average age when the 50 biological scientists who received the AACS fellowship was approximately 50 years, with a wide age range of 29–80 years. The average age of the scientists receiving the NMS was approximately 69 years, ranging between 55 and 90 years. This finding implied that receiving the NMS is a higher science achievement than obtaining the AACS fellowship is. Most of the biological scientists were bestowed the NMS at the end of their academic career, thus indicating that those scientists have contributed their lives to research before their contributions were acknowledged. The time length between receiving the AACS fellowship and the NMS ranged between 1 and 40 years. Only two scientists in the study sample received NMS before obtaining the AACS fellowship, whereas the other 48 scientists received it on an average of approximately 20 years after obtaining the AACS fellowship. Moreover, they were affiliated with 30 institutions. The Massachusetts Institute of Technology was the institution with the highest number of scientists (six scientists).

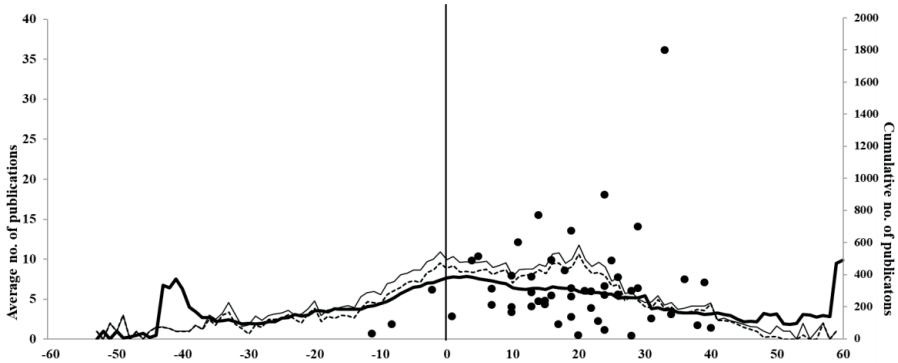
Table 1. List of Scientists

No	Name	Institution	Age (NMS)	Age (AACS)	Ave. Publications*	Ave. research influence**
1	Robert A. Weinberg	Whitehead Institute for Biomedical Research; Massachusetts Institute of Technology	55	45	6.78	437.38
2	Elaine Fuchs	Howard Hughes Medical Institute; Rockefeller Univ.	58	44	8.43	156.79
3	Francis S. Collins	Natl. Institutes of Health	58	48	10.94	244.44
4	May Berenbaum	Univ. of Illinois Urbana-Champaign	59	43	8.00	49.70
5	Ann M. Graybiel	Massachusetts Institute of Technology	59	49	5.29	123.12
6	Susan Lee Lindquist	Massachusetts Institute of Technology	60	47	7.92	175.56
7	Phillip A. Sharp	Massachusetts Institute of Technology	60	39	7.78	203.64
8	David Baltimore	California Institute of Technology	61	36	11.02	216.87
9	Lynn Margulis	Univ. of Massachusetts	61	60	3.74	26.03
10	J. Craig Venter	J. Craig Venter Institute	62	55	5.76	333.77
11	Harold Varmus	Memorial Sloan-Kettering Cancer Center	62	49	6.18	165.74
12	Nancy C. Andreasen	Univ. of Iowa	62	64	10.93	116.22
13	Jared Diamond	Univ. of California, Los Angeles	62	36	5.25	74.62
14	Rakesh K. Jain	Harvard Medical School; Massachusetts General Hospital	63	58	13.41	160.58
15	Sallie W. Chisholm	Massachusetts Institute of Technology	64	45	3.73	137.10
16	Robert J. Lefkowitz	Howard Hughes Medical Institute; Duke Univ.	64	45	15.33	167.32
17	Nina V. Fedoroff	Pennsylvania State Univ.	64	47	2.47	71.23
18	Mario R. Capecchi	Univ. of Utah	64	72	4.06	128.46
19	Peter H. Raven	Missouri Botanical Garden and Washington Univ.	64	41	2.35	40.86
20	Anthony S. Fauci	Natl. Institutes of Health	65	51	16.87	138.38
21	Solomon H. Snyder	Johns Hopkins Univ.	65	41	18.24	183.19
22	Gene E. Likens	Institute of Ecosystem Studies	66	44	5.60	95.34
23	Stanley Prusiner	Univ. of California, San Francisco	67	51	10.58	141.88

24	J. Michael Bishop	Univ. of California, San Francisco	67	48	6.69	160.07
25	Francisco J. Ayala	Univ. of California, Irvine	67	43	7.70	45.91
26	Mary-Claire King	Univ. of Washington	68	53	4.84	124.32
27	Rudolf Jaenisch	Whitehead Institute for Biomedical Research	68	50	9.98	220.62
28	Lubert Stryer	Stanford Univ.	68	37	2.39	191.33
29	George F. Bass	Texas A&M Univ.	69	80	0.64	8.53
30	James D. Watson	Cold Spring Harbor Laboratory	69	29	1.24	285.30
31	Bruce N. Ames	Univ. of California, Berkeley	70	42	6.52	273.55
32	Lucy Shapiro	Stanford Univ.	71	52	5.40	72.06
33	Bert W. O'Malley	Baylor College of Medicine	71	60	13.52	98.50
34	Alexander Rich	Massachusetts Institute of Technology	71	35	6.95	104.31
35	Rita R. Colwell	Univ. of Maryland	72	68	11.10	55.27
36	James E. Darnell	Rockefeller Univ.	72	43	5.31	209.62
37	Carl R. Woese	Univ. of Illinois	72	57	4.00	191.30
38	Simon A. Levin	Princeton Univ.	73	51	6.09	98.59
39	Leroy Hood	Institute for Systems Biology	73	44	14.57	135.62
40	Janet D. Rowley	Univ. of Chicago	73	66	7.65	97.27
41	Bruce Alberts	Univ. of California, San Francisco	74	40	2.75	126.24
42	Ralph Brinster	Univ. of Pennsylvania	78	54	6.09	137.34
43	Thomas E. Starzl	Univ. of Pittsburgh	78	45	27.60	53.77
44	Charles Yanofsky	Stanford Univ.	78	39	5.76	65.96
45	Stanley Falkow	Stanford Univ.	80	54	6.64	137.48
46	Victor A. McKusick	Johns Hopkins Univ.	80	67	6.73	90.66
47	Torsten N.Wiesel	Rockefeller Univ.	81	43	1.50	563.40
48	Evelyn M. Witkin	Rutgers Univ.	81	57	0.84	115.92
49	Ruth Patrick	Academy of Natural Sciences	89	69	0.46	11.05
50	Norman E. Borlaug	Texas A&M Univ.	90	62	0.48	18.63

Note: * refers to the annual average number of three types of publications covered by WoS and Scopus; ** refers to the average number of citations received by each of three types of publications covered by only WoS.

The number of publications (articles, conference papers, and review articles) covered by WoS and Scopus published by each scientist ranged between 31 and 1949, which were shown with the height of the dots based on the legend on the right vertical axis in Figure 1. The annual average number of publications for each scientist ranged between 0.46 and 27.60 (Table 1). Figure 1 shows the marginal gap between the normal solid line and the dotted line, indicating that the 50 biological scientists examined in the study published most of the three types of publications in journals covered by WoS. An average of approximately 89.2% of publications per scientist were articles, ranging between 71.2% and 100.0%. Moreover, an average of 84.8% of publications for each scientist were covered by WoS. The percentage of publications indexed by Scopus and those not covered by WoS for each scientist ranged between 3.2% and 38.4%. Over 25% of publications by 11 scientists were not covered by WoS.



Notes: --- refers to the annual average number of publications from only WoS for each scientist; ─ refers to the annual average number of publications covered by WoS and Scopus for each scientist; ▬ refers to the annual average number of citations per publication for each scientist, according to the WoS.

Figure 1. Annual average research productivity and research influence for each scientist

Figure 1 shows the annual average number of publications for each scientist, indicating that the research productivity peaked around the year when the scientists obtained the AACS fellowship, which is labeled as zero in the horizontal axis of a graph plot that represents the time length. The dots located in the horizontal axis refer to the year when a specific scientist received the NMS. This indicates that only three scientists received the NMS before the AACS fellowship. Moreover, with the exception of two ends of the bold solid line, similar trends were noted in the annual average number between research productivity (the normal solid line) and research influence (the solid bold line). However, the average number could not indicate the real research performance for the 50 scientists in the study because of significant differences in research performance at individual levels.

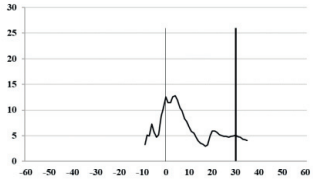
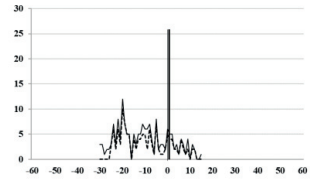
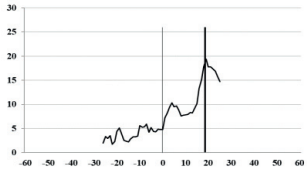
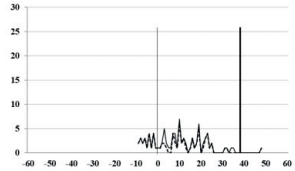
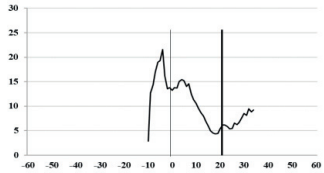
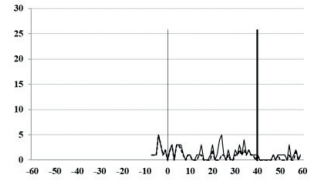
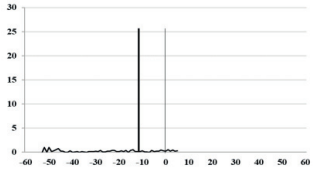
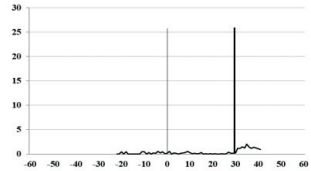
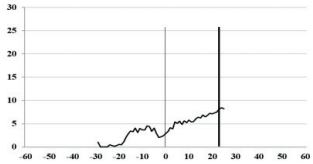
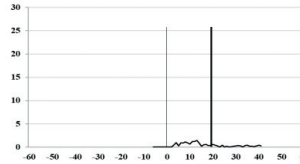
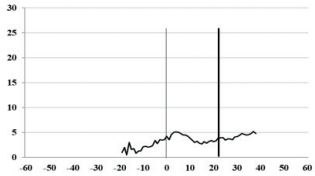
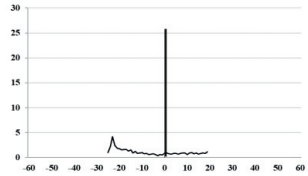
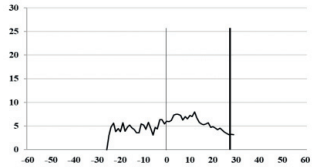
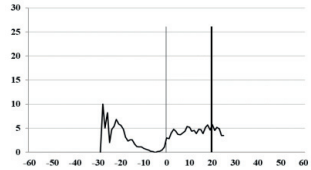
Types of research productivity

Based on the changes in the annual number of publications in three periods, 50 biological scientists were classified into 10 types of research productivity (RP). In the examples provided in Table 2, the year when each scientist obtained the AACS fellowship was labeled as zero in the horizontal axis and a gray vertical line was marked at the zero point. The point where a black vertical line was marked referred to the year when a specific scientist received the NMS. Table 2 shows that types 112 and 122 were indicated for most scientists in terms of research productivity, with 15 scientists indicating these types. An increasing trend was noted in the first and second period for Type 112, whereas a decreasing trend was noted in the third period (after receiving NMS). An increasing trend was noted in the first period for Type 122, whereas a decreasing trend was noted the other two periods. Moreover, three scientists indicated low research productivity, with no trends revealed in three periods (Type 000).

Types of research influence

A substantial difference in the mean annual number of citations received per publication was noted among the scientists, with a wide range of 8.5 and 563.4 (Table 1). To examine the changes in the mean annual number of citations received per publication within the three periods, 11 types of research influence (RI) were observed. As shown in Table 2, 23 scientists (46%) belonged to type 122. Type 121 was the second most common type (8 scientists, 16%), followed by type 112 (6 scientists, 12%).

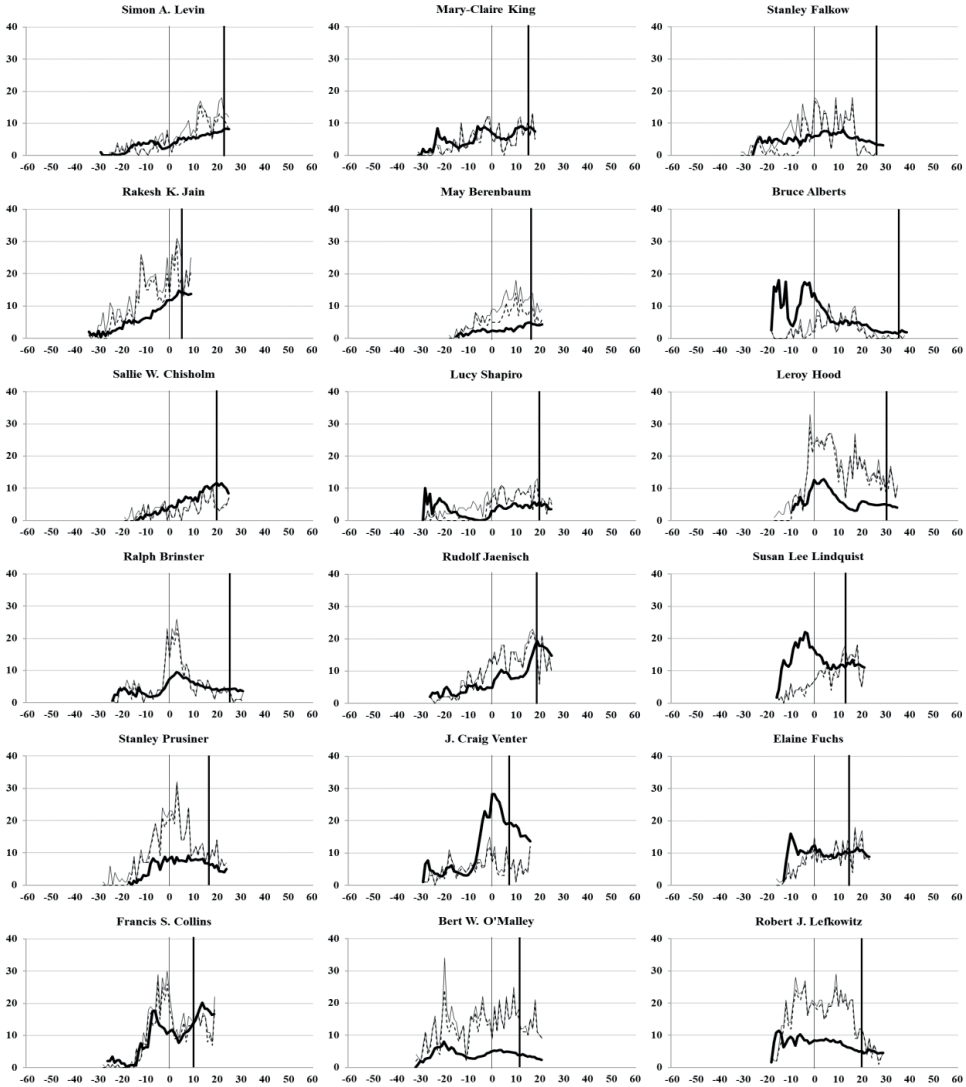
Table 2. Types of research productivity and research influence

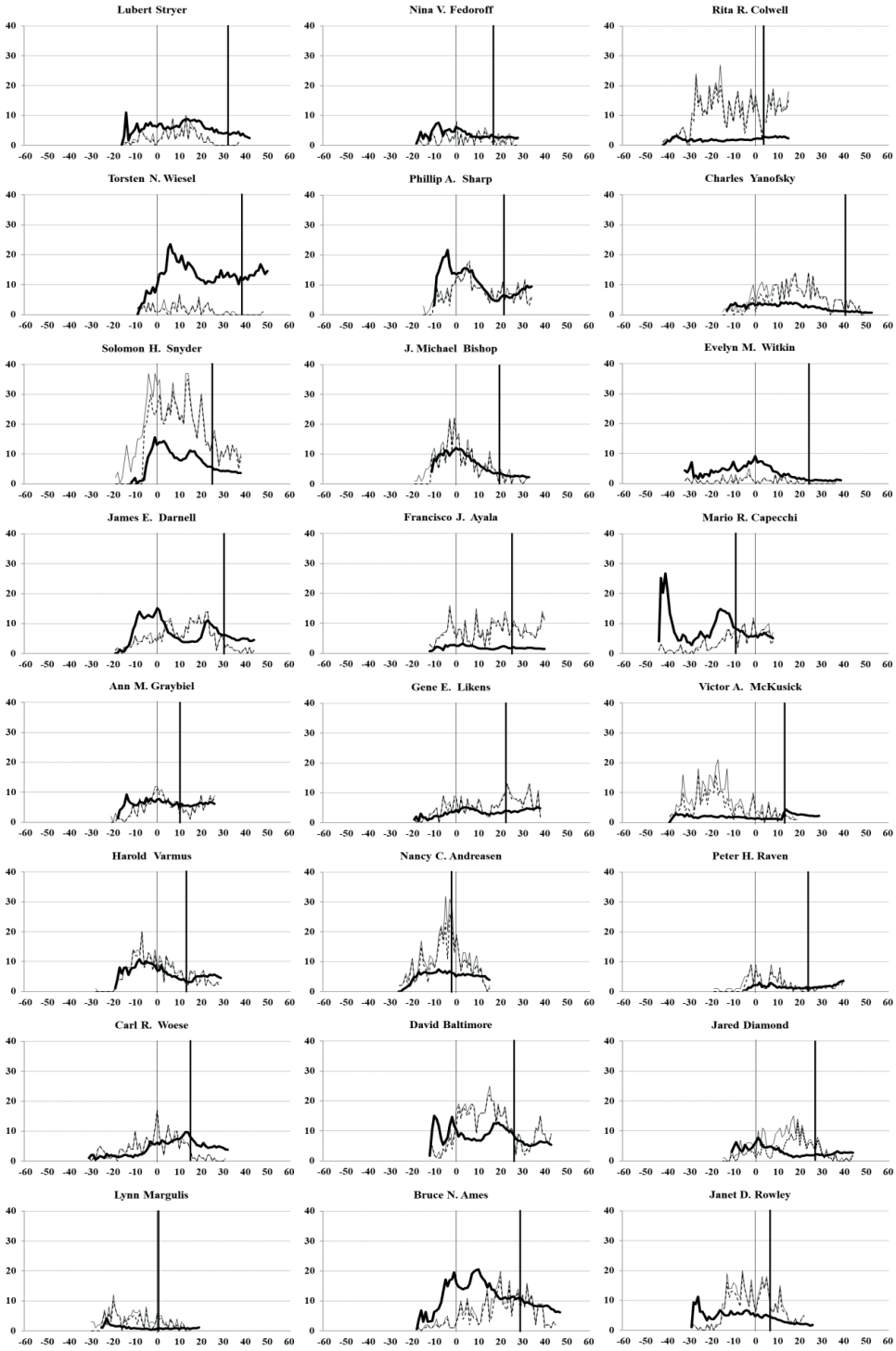
Type	Example	RP	RI	Type	Example	RP	RI
122		15	23	102		1	0
112		15	6	220		1	0
121		5	8	221		1	0
000		3	1	001		0	1
110		3	2	010		0	1
111		3	3	200		0	1
120		3	2	212		0	2

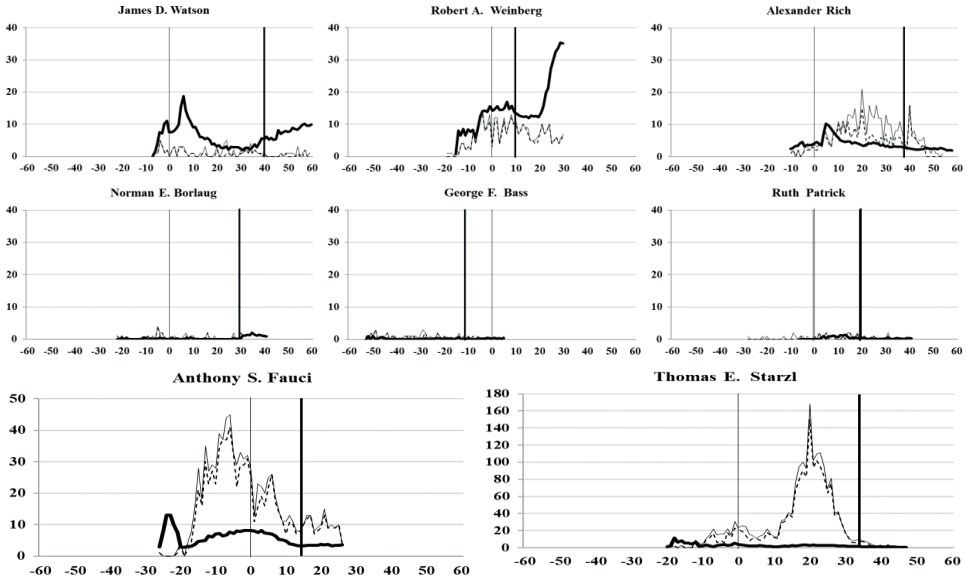
Comparison of research productivity and research influence

Figure 2 shows that seven common types existed between research productivity and research influence. The types of research productivity and that of research influence were not similar for some scientists. Differences in the types between research productivity and research influence for each scientist were used to classify 50 scientists into seven broad types of research performance. The types of research productivity and that of research influence for each scientist are shown in Figure 2. wherein the bold solid line refers to research influence, the solid line refers to research productivity limited to publications covered by WoS, and the dotted line refers to research productivity based on publications indexed by WoS and Scopus.

Figure 2. Research performances for individual scientists




Lubert Stryer
Nina V. Fedoroff
Rita R. Colwell
Torsten N. Wiesel
Phillip A. Sharp
Charles Yanofsky
Solomon H. Snyder
J. Michael Bishop
Evelyn M. Witkin
James E. Darnell
Francisco J. Ayala
Mario R. Capecchi
Ann M. Graybiel
Gene E. Likens
Victor A. McKusick
Harold Varmus
Nancy C. Andreasen
Peter H. Raven
Carl R. Woese
David Baltimore
Jared Diamond
Lynn Margulis
Bruce N. Ames
Janet D. Rowley



Notes: Bold lines refer to research influence; normal lines refer to research productivity based on WoS and the Scopus; dotted lines refer to research productivity based on only WoS.

To provide a snapshot of the characteristics of long-term research performance of 50 scientists, 50 scientists were divided into seven types. Table 3 shows that type A was the primary type with 19 scientists and features the same six types for both research productivity and research influence. Type B was the second most common type with 12 scientists. A similar trend was noted in the first and third periods in both types of research productivity and research influence. Type D (8 scientists) came in third, with similar trends noted in the first two periods, followed by type C (six scientists), with an increasing trend in the first period in both research productivity and research influence, but an opposite trend in the second and third periods. Two scientists were indicated in the types E and G. For type E, a different trend was noted in the first period, whereas a similar trend was noted in the second and third periods. For type G, a similar trend was noted in the second period, whereas a different trend was observed in the first and third periods. Only one scientist was classified as type F; a different trend was noted in the first two periods, whereas a similar trend was noted in the third period.

Table 3. Types of research performance

Type	No. of scientists	Pair of types of research productivity and research influence
A	19	(000,000), (110,110), (112,112), (120,120), (121,121), (122,122)
B	12	(000,010), (112,122), (121,111), (122,112)
C	6	(111,122), (112,121), (112,121), (121,112), (122, 111)
D	8	(000,001), (111,112), (112,111), (120,122), (121,122), (122,121)
E	2	(112,212), (221,121)
F	1	(122,212)
G	2	(102,200), (220,121)
Total	50	

Discussion and Conclusion

The results of this study confirm that 50 notable biological scientists with significant contributions to the field did not indicate the same type of research performance. The age of the 50 scientists ranged between 65 and 105 years in 2018, whereas 11 scientists have already passed away. Moreover, this study collected their publications before 2018. Therefore, the findings of this study represent the research performance of most of the scientists' entire academic career. The results are not consistent with those of Chan et al. (2014), which indicated that increased research productivity and research influence were observed among scientists after they received notable scientific awards. The difference in the study results may be because the scientists were awarded the NMS during their later years. An inverted U-shaped relationship existed between their research age and productivity (Perlin, Santos, Imasato, Borenstein, & Da Silva, 2017). As expected, the scientists were not as prolific during the late period of their academic careers. The status of the NMS is higher than that of the AACS fellowship. Therefore, scientists were awarded the NMS after they had obtained the AACS fellowship, although three scientists are exceptions. Most of the scientists (26 scientists) achieved the peak of their research productivity during the second period, after they had received the AACS fellowship, and before they received the NMS. Moreover, 20 scientists achieved the peak of their research productivity before they obtained the AACS fellowship. Only four scientists were found to have the highest research productivity after they had received the NMS.

Regarding the annual average research influence per publication, this did not reveal the real research influence for most scientists because of the skewness of citations of publications. Some NMS scientists were found to have no substantial research influence over the course of their academic career and had indicated decreasing trends in their research influence during the second period before they had received the NMS. After examining their research influence for each publication, with the exception of four scientists, 26 other scientists had published at least one article with over 500 cumulative citations until 2017. Two scientists each had an article with more than 10,000 cumulative citations. The annual average citations for the publication with the highest cumulative number of citations among the 26 scientists ranged between 15 and 2,782. This finding reveals that most biological scientists who have received the NMS are highly influential researchers.

Among four scientists who indicated low research productivity (annual number of publications lower than 4) and low research influence (publications with less than 500 citations), two scientists were found to have influential books, based on a substantial number of citations obtained from the reference search function provided by WoS. However, books are not one of the primary types of research that are indexed by WoS and Scopus. Therefore, the research performance of scientists who published books were underestimated. Biological scientists tend to publish journal articles (Bourke & Butler, 1996; Mutz, Bornmann, & Daniel, 2013). Most researchers in the fields of science and technology demonstrate their research results and influence through research articles. However, few scientists' research performance may be primarily published through books. Therefore, books must be considered for measuring the research performance of scientists. Moreover, differences continue to exist between the subfields within biology. One biological scientist may specifically focus on underwater archaeology, a rare and small profession. This may lead to his books having no substantial citations, and therefore, be incomparable with the research influence of other scientists. Another biological scientist may alleviate the human hunger problem, and therefore, may have contributed substantially to help people around the world. However, this type of contribution may not be sufficiently reflected in their research publications. Research results not only serve the scientific community. Improving the quality of life and resolving social problems are goals

of scientific research. Therefore, some aspects of research performance are not measured using bibliometric indicators related to research performance.

To summarize the study findings, research productivity has a weaker association with scientific contributions than research influence does. The substantial scientific contributions of notable biological scientists with NMS can be proved based on their highly cited publications. Although few publications are influential for most scientists, they have been cited a substantial number of times for a long period. Therefore, the average citation per publication is not an appropriate indicator for identifying excellent scientists. Various types of research productivity and research influence cannot indicate a positive association between research performances and awards for each scientist. To enhance the precision of examining research performance, books and citations received for books must also be added to the measurements.

Acknowledgements This work was financially supported by the Center for Research in Econometric Theory and Applications (Grant no. 108L900204) from The Featured Areas Research Center Program within the framework of the Higher Education Sprout Project by the Ministry of Education (MOE) in Taiwan, and by the Ministry of Science and Technology (MOST), Taiwan, under Grant No. MOST 108-3017-F-002-003- and MOST 107-2410-H-002-200-

References

Abramo, G., D'angelo, C. A., & Solazzi, M. (2011). The relationship between scientists' research performance and the degree of interNatl.ization of their research. *Scientometrics, 86*(3), 629-643.

Borjas, G. J., &Doran, K. (2015). Prizes and productivity: How winning the fields medal affects scientific output. *Journal of Human Resources, 50*(3), 728-758.

Bourke, P., & Butler, L. (1996). Publication types, citation rates and evaluation. *Scientometrics, 37*(3), 473-494.

Chan, H. F., Frey, B. S., Gallus, J., & Torgler, B. (2014). Academic honors and performance. *Labour Economics, 31*, 188-204.

Khosravi, A., & Chavan, M. (2012). A comprehensive view on research excellence. *Total Quality Management, 23*(5), 507-521.

Licea De Arenas, J., Valles, J., Arenas, M. (1999). Profile of the Mexican health sciences elite: A bibliometric analysis of research performance. *Scientometrics, 46*(3), 539-547.

Mutz, R., Bornmann, L., & Daniel, H. D. (2013). Types of research output profiles: A multilevel latent class analysis of the Austrian Science Fund's final project report data. *Research Evaluation, 22*(2), 118-133.

National Science Foundation (2018). Nomination information. Retrieved April 7, 2019 from: https://www.nsf.gov/od/nms/nominations.jsp

O'Connell, D., & Rugman, A. (2013). The research productivity of the AIB Farmer Award finalists. Multinaitonal Business Review, 21(4), 298-311.

Perlin, M. S., Santos, A. A. P., Imasato, T., Borenstein, D., & Da Silva, S (2017).The Brazilian scientific output published in journals: A study based on a large CV database. *Journal of Informetrics, 11*(1), 18-31.

Rodríguez-Navarro, A. (2011). Measuring research excellence: Number of Nobel Prize achievements versus conventional bibliometric indicators. *Journal of Documentation, 67*(4), 582-600.

Yair, G., Gueta, No., & Davidovitch, N. (2017). The law of limited excellence: Publication productivity of Israel Prize laureates in the life and exact sciences. *Scientometrics, 113*, 299-311.

Bibliographically Coupled Patents: Their Temporal Pattern and Combined Relevance

Chung-Huei Kuan[1,3], Dar-Zen Chen[2,3]

[1] *maxkuan@mail.ntust.edu.tw*
National Taiwan University of Science and Technology, Graduate Institute of Patent, No. 43, Sec. 4, Keelung Rd., Taipei, Taiwan (R.O.C.)

[2] *dzchen@ntu.edu.tw*
National Taiwan University, Department and Graduate Institute of Mechanical Engineering and Graduate Institute of Industrial Engineering, No. 1, Sec. 4, Roosevelt Rd., Taipei, Taiwan (R.O.C)

[3] National Taiwan University, Center for Research in Econometric Theory and Applications (CRETA), No. 1, Sec. 4, Roosevelt Rd., Taipei, Taiwan (R.O.C.)

Abstract

Bibliographic coupling (BC) is one of the most common indicators in detecting and measuring patent relatedness. Patents that are bibliographically coupled reveal a temporal pattern involving their ages (how long ago they are issued) and spans (their distances in time). BC is more frequently found between patents issued more recently and closer in time, and their coupling strength also tends to be stronger. Aged or long-spanned patent pairs are not only fewer but also inherently limited in their coupling strength. These patent pairs therefore may be overlooked when a threshold is applied, even though their patents are highly related. An improved measure, referred to as *combined relevance*, is proposed to provide fairer treatment to these aged or long-spanned patent pairs when assessing the relatedness between their patents. Combined relevance is as simple as the conventional measures, both conceptually and computationally. More importantly, a fixed threshold may be safely applied with a reduced possibility of erroneously removing the aged or long-spanned patent pairs.

Introduction

A lot of patent bibliometric works involve the detection and measurement of relatedness between patents. Based on their relatedness, structures may be derived from thousands of seemingly disorganized patents. Then, collections of related patents may be modelled and monitored as a unit; evolving trends may be detected by observing related patents or patent clusters in their chronological orders. The cooperation/competition relationship and knowledge exchange between firms, institutions, counties, and fields may be examined and inferred based on the relatedness between their patents.

There are three major categories of approaches in detecting and measuring patent relatedness. The text-based approaches extract textual information from patents' specifications (cf. Arts, Cassiman, & Gomez, 2018; Moehrle & Gerken, 2012; Niemann, Moehrle, & Frischkorn, 2017; Ortiz-de-Urbina-Criado, Nájera-Sánchez, & Mora-Valentín, 2018; Yoon & Magee, 2018). The classification-based approaches evaluate the overlapping of classification symbols assigned to patents (cf. Angue, Ayerbe, & Mitkova, 2014; Chang, 2012; Jaff, 1986; Kuan et al., 2018; Petruzzelli, 2011; Wang, Hou, & Hung, 2018). The third category, citation-based approaches, generally involve three common indicators: direct citation (DC) (Trajtenberg, 1990), bibliographic coupling (BC) (Kessler, 1963), and co-citation (CC) (Small, 1973). These mechanisms may be used individually or together. Kuusi and Meyer (2007) employed BC alone to cluster some related patents and identify an emerging technological paradigm. Lo (2007) also employed only BC to identify technological connections between major research organizations. Von Wartburg, Teichert, and Rost (2005) combined DC and BC in a multistage analysis to reveal technological change. Chen, Huang, Hsieh, and Lin (2011), Yeh, Sung, Yang, Tsai, and

Chen (2013), and Kuan, Huang, and Chen (2018) used both DC and BC to construct more comprehensive citation networks. Citation-based approaches may also be applied together with the other two approaches. Leydesdorff, Kushnir, and Rafols (2014) and Kuan et al. (2018) integrated DC with patent classification codes, Nakamura, Suzuki, Sakata, and Kajikawa (2015) combined DC and co-word analysis, and Park, Jeong, Yoon, and Mortara (2015) used BC and patent text semantic analysis to locate potential R&D collaboration partners.

DC, BC, and CC are all valid relatedness indicators, as evidenced by a large body of works where the aforementioned are only a few samples. These indicators, however, may conflict when one suggests relatedness whereas another specifies otherwise. In a previous study investigating one such conflict involving patents that do not cite each other but are strongly bibliographically coupled, the authors found that this phenomenon is not coincidental (Kuan et al., 2018). For DC to occur, the cited patent has to be published earlier so that it is visible and citable to the applicant or examiner of the citing patent. Therefore, DC rarely occurs between patents whose application processes are highly overlapped, as their applicants or examiners are blind to each other, even though the two patents are indeed related. BC is not handicapped as such and may effectively reveal the patents' relatedness.

During the previous study, the authors also noticed that bibliographically coupled patents reveal a temporal pattern involving their ages (how long ago they have been issued) and spans (their distances in time) and this pattern would place an inherent bound on their coupling strength. In the following sections, this pattern, its cause, and its implication on the traditional measurement and relatedness assessment are described and discussed. Then, an improved measure is proposed to reduce the impact of ages and spans in assessing bibliographically coupled patents.

Temporal Pattern and Cause

A. Phenomenon

To demonstrate this temporal pattern, the same dataset from the previous study is borrowed here, which includes 34,083 US utility patents issued between 1976/1/1 and 2017/3/31 in the field of carbon dioxide capture, storage, recovery, delivery, and regeneration. Among the 34,083 patents, there are 154,505 pairs of cited and citing patents, 1,609,549 pairs of bibliographically coupled patents, and 644,376 pairs of co-cited patents, hereinafter respectively referred to as DC pairs, BC pairs, and CC pairs. Each of the DC, BC, and CC pairs includes patents P_E and P_L respectively issued at an earlier date t_E and a later date t_L ($t_E \leq t_L$). Then, the *span* and *age* of a DC, BC, or CC pair are respectively defined as $t_L - t_E$ and $t_{NOW} - t_E$, where t_{NOW} denotes the cut-off date of the patent data collection (e.g., 2017/03/31 for this dataset).

Figure 1 shows the frequency distributions of all DC, BC, and CC pairs according to their ages (horizontal axis) and spans (vertical axis) in years, where more reddish or bluish points reflect higher or lower counts. BC pairs, in contrast to DC and CC pairs, are particularly concentrated in the lower left corner, meaning significantly more BC pairs have shorter spans and smaller ages, or BC is more frequently found between patents issued more recently and closer in time. DC and CC do not reveal such propensity.

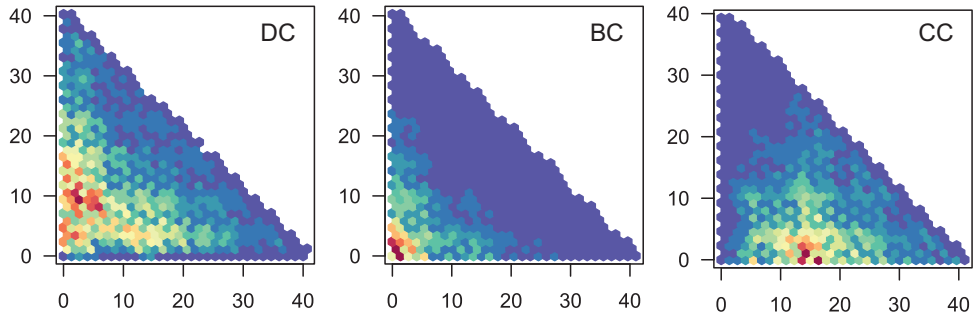


Figure 1. Frequency distributions of DC (left), BC (middle), and CC (right) pairs according to ages (x) and spans (y).

Not only that, the bibliographic coupling strength (BCS) of the BC pairs also reveals a similar pattern. Figure 2 shows the average BCS and average co-citation strength (CCS) for BC and CC pairs across ages and spans, where BCS and CCS are measured as the number of cited and citing patents in common. Again, BC pairs having higher BCS (i.e., more reddish points) are particularly concentrated in the lower left corner, whereas CCS does not show any significant pattern.

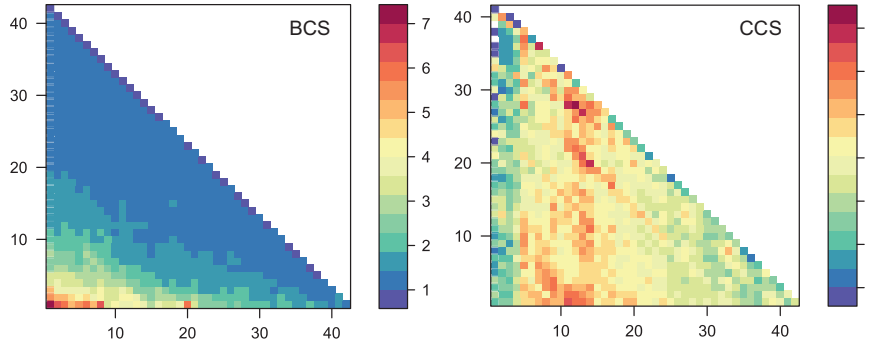


Figure 2. Average BCS (left) and average CCS (right) according to ages (x) and spans (y).

Swanson (1971) and Jarneving (2007a) indicated that only BC pairs having BCS above a threshold are truly related. Indeed, among the 1,609,549 BC pairs, up to 1,167,794 (72.55%) of them have the smallest BCS of 1. To avoid that the pattern shown in Figure 2 is resulted from a large volume of noises, Figure 3 shows the distributions of average BCS after removing those BC pairs having BCS not greater than 1 and the overall average BCS (2.74 or μ). The same pattern is still preserved.

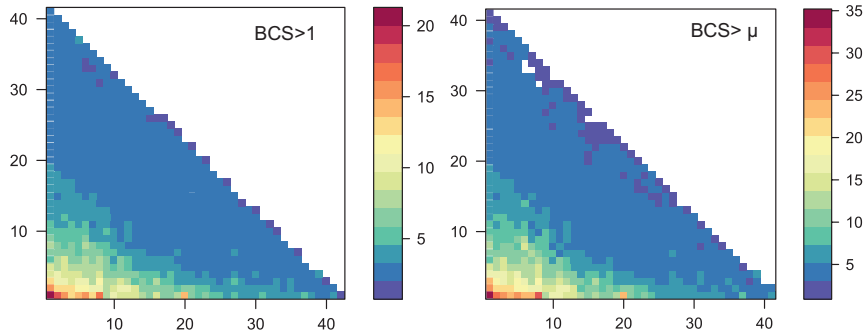


Figure 3. Average BCS for pairs with BCS > 1 (left) and BCS > μ (right) according to ages (x) and spans (y).

B. Patent and Reference Expansion

This unique pattern is related to a field's continuously increasing numbers of accumulated patents and cited references, referred to as *citable patent expansion* and *cited reference expansion* subsequently. Figure 4 depicts the two expansions using the same case data. The horizontal axes show the ages of patents in 2-year intervals. For patents issued earlier in the past or more recently, their data are plotted more to the right or to the left. In the left diagram, the bars show the numbers of patents issued within respective intervals relative to the left scale; the curve depicts the numbers of accumulative patents or citable patents up to each interval. The curve in the right diagram shows the average numbers of cited references at respective intervals.

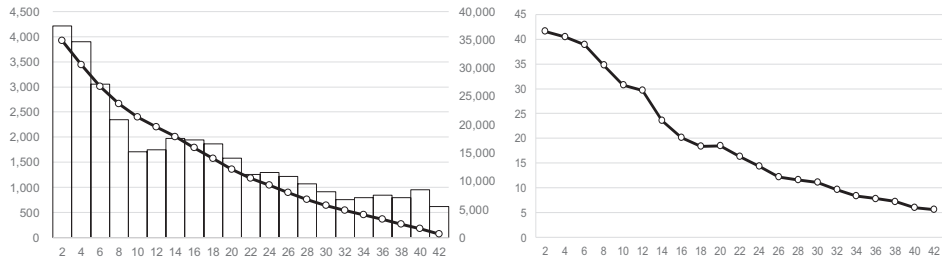


Figure 4. A field's expanding numbers of citable patents (left) and cited references (right).

As illustrated in the left diagram, the citable patent curve rises monotonically from right to left as patents of the field are issued and accumulated over time. Then, later patents have more citable patents and, therefore, may have more references than earlier patents do. The average cited reference curve of the right diagram, as such, also rises monotonically from right to left.

The citable patent expansion should be applicable to patents from any field and from any patent office. The universal applicability of cited reference expansion is, however, questionable. Researchers did notice that the number of references made per U.S. patent and the number of U.S. patents issued both increase over time (Hall, Jaffe, & Trajtenberg, 2001; Zhang, Huang, & Chen, 2018). However, the authors speculates that cited reference expansion should hold for U.S. patents, as U.S. regulation obligates applicants to disclose (i.e., cite) all information (e.g.,

prior patents) known to be relevant to the applications' patentability (Bicknell, 2008). U.S. applicants therefore tend to cite more when there are more citable patents.

Facing the citable patent and cited reference expansions, Hall, Jaffe, and Trajtenberg (2001) and Zhang, Huang, and Chen (2018) were concerned about the "devaluation" of citations (i.e., a patent's "later citations are less significant than earlier ones"). This study, however, is concerned about their implication on BC and patent relatedness based on BC.

Implication on Conventional Assessment

A. Inherent Bound

Figure 5 depicts four scenarios between patents P_E and P_L from a same field, where the bars are the numbers of patents issued within respective intervals from the left diagram of Figure 4. P_E and P_L have references REF_E and REF_L respectively drawn from citable patents CP_{t_E} and CP_{t_L} accumulated up their issue dates t_E and t_L. P_E and P_L should satisfy the following Eqs. (1) and (2):

$$REF_E \subseteq CP_{t_E}, \quad |REF_E| \leq |CP_{t_E}|, \text{ and} \tag{1}$$

$$REF_L \subseteq CP_{t_L}, \quad |REF_L| \leq |CP_{t_L}|. \tag{2}$$

Then, Eq. (3) may be derived according to the citable patent expansion:

$$CP_{t_E} \subseteq CP_{t_L}, \quad |CP_{t_E}| \leq |CP_{t_L}|. \tag{3}$$

CP_{t_E} is denoted by dark grey bars, and CPP_{t_L} is denoted by both dark and light grey bars in Figure 5. If the cited reference expansion is applicable, then

$$|REF_E| \leq |REF_L|, \text{ and} \tag{4}$$

$$|REF_E \cap REF_L| \leq \min(|REF_E|, |REF_L|) = |REF_E| \leq |CP_{t_E}|. \tag{5}$$

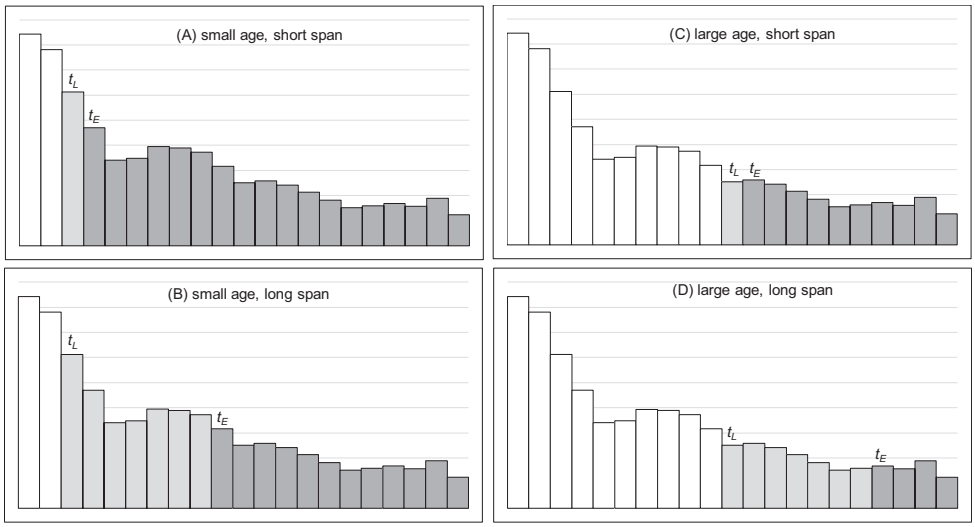


Figure 5. Four scenarios between patents P_E and P_L from a same field.

For P_E and P_L to form a BC pair, they need to have non-empty intersection or $REF_E \cap REF_L \neq \emptyset$. Under cited reference expansion, $|REF_E \cap REF_L|$ cannot be greater than $|REF_E|$, which in turn is bounded $|CP_{t_E}|$. CP_{t_E} is the set of CPs common to both P_E and P_L denoted by the dark

grey bars. Therefore, larger CP_{t_E} implies a greater chance in achieving non-empty intersection and thus forming a BC pair. For the four scenarios, whether t_L is more recent as in diagrams (A) and (B) or earlier in the past as in diagrams (C) and (D), $|CP_{t_E}|$ is greater when P_E and P_L have a shorter span. On the other hand, whether P_E and P_L have a shorter span in diagrams (A) and (C) or a longer span as in diagrams (B) and (D), $|CP_{t_E}|$ is greater when t_L is more recent. This is why more BC pairs have shorter spans and smaller ages as illustrated in Figure 1. Similarly, for a BC pair involving P_E and P_L, its BCS, $|REF_E \cap REF_L|$, is also bounded by $|CPP_{t_E}|$, and larger CPP_{t_E} provides a greater chance in achieving higher BCS. This is why BC pairs having shorter spans and smaller ages tend to have greater BCS as illustrated in Figures 2 and 3.

For BC pairs to deliver the unique temporal behavior, the cited reference expansion or Eq. (4) should hold. This is indeed true for a major portion of the case's BC pairs. Table 1 lists the shares of BC pairs satisfying Eq. (4), $|REF_E| \leq |REF_L|$, among those whose BCS crosses ten different thresholds. For all BC pairs (BCS>0), despite a large number of BC pairs that may be noises, there is still more than 65% of the BC pairs satisfying $|REF_E| \leq |REF_L|$. At a higher threshold, there are even greater portions satisfying $|REF_E| \leq |REF_L|$.

Table 1. Shares of BC pairs satisfying $|REF_E| \leq |REF_L|$.

BCS>	*0*	*2*	*4*	*6*	*8*	*10*	*12*	*14*	*16*	*18*
All BC pairs	1,609,549	214,324	94,090	59,260	43,381	34,989	29,357	25,882	23,283	21,284
% of $\|REF_E\| \leq \|REF_L\|$	65.46%	66.44%	67.70%	68.46%	69.54%	70.62%	71.70%	72.49%	73.16%	73.59%

B. Problem with Conventional Measure and Threshold

Using a threshold to filter BC pairs, despite a common practice, may not be appropriate for a field revealing the temporal pattern. For example, if a threshold for BCS is set to the overall average (2.74) of the case's all BC pairs, pretty much the BC pairs denoted by the greenish or bluish points in Figure 2 are filtered out, even though some of these aged or long-spanned BC pairs may reflect true relatedness.

This filtering-without-distinction problem not only due to the fixed threshold use, but also due to the BCS measure's overlooking the age and span factors. The frequently used BCS measures may be classified into two broad categories: *intersection-based* and *vector-based* measures with Jaccard coefficient (Jaccard, 1901) and coupling angle (Glänzel, & Czerwon, 1996) as representatives. If Jaccard coefficient or coupling angle is applied to a field with the temporal pattern, all BCS would be bounded by $|REF_E|$, the size of the earlier patent P_E, as derived from the following Eqs (6) and (7):

$$\frac{|REF_E \cap REF_L|}{|REF_E \cup REF_L|} \leq \frac{|REF_E|}{|REF_E \cup REF_L|} \leq |REF_E| \text{ , and} \tag{6}$$

$$\frac{\overrightarrow{REF_E} \cdot \overrightarrow{REF_L}}{|\overrightarrow{REF_E}||\overrightarrow{REF_L}|} = \frac{|REF_E \cap REF_L|}{|\overrightarrow{REF_E}||\overrightarrow{REF_L}|} \leq \frac{|REF_E|}{|\overrightarrow{REF_E}||\overrightarrow{REF_L}|} \leq |REF_E| \text{ ,} \tag{7}$$

where $\overrightarrow{REF_E}$ and $\overrightarrow{REF_L}$ are REF_E and REF_L expressed in binary vectors of equal dimension.

Bibliometric researchers had long noticed the age and span problem. For example, Jarneving (2007b) indicated that "an increase of the distance in time between bibliographically coupled articles leads to a diminishing pool of shared references as there is a tendency to cite the more

current articles" That is why usually an observation window is set up so that bibliographically coupled research articles published closer (i.e., about the same age) within the window (i.e., within limited span) are collected and compared together (cf. Jarneving, 2007b; Glänzel, & Czerwon, 1996).

Combined Relevance

However, to observe the knowledge flow or to develop a representative trajectory among patents across a long period of time, where all relevant BC pairs have to be considered, a BCS measure as much immune to their ages and spans as possible would be desirable.

This study therefore proposes a new and simple BCS measure, referred to as *combined relevance* (CR), in Eq. (8):

$$\left(\frac{|REF_E \cap REF_L|}{|REF_E|}\right)\left(\frac{|REF_E \cap REF_L|}{|REF_L|}\right) = \frac{|REF_E \cap REF_L|^2}{|REF_E||REF_L|}. \quad (8)$$

The idea behind Eq. (8) is straightforward. Imaging that REF_E and REF_L respectively represent the information gathered by a BC pair's earlier and later patents P_E and P_L, and that $REF_E \cap REF_L$ is the piece of information shared between them, the two factors in Eq. (8) measure how much this shared information accounts for P_E and P_L's gathered information, or this shared information's *individual relevance* to P_E and P_L. Then, P_E and P_L are considered highly related if their shared information is relevant to both of them.

Figure 6 shows the averages of the two factors for BC pairs whose BCS is greater than 1 in two separate diagrams. The BC pairs are limited to those having BCS > 1 so that the observation is not impaired by a large volume of noises. As illustrated in the left diagram, P_E's factor tends to have higher values for those aged or long-spanned BC pairs located farther away from the lower left corner. This is because their P_E occurs earlier in the field and REF_E is more limited (see Figure 4, right diagram). Then, P_E's factor would be close to 1 as $|REF_E \cap REF_L| \sim |REF_E|$. On the other hand, P_L's factor tends to have higher value for those aged but short-spanned BC pairs located closer to the lower right corner. This is because their P_L also occurs earlier in the field, and REF_L is not only more limited but also close to REF_E, therefore P_L's factor would be close to 1 as $|REF_E \cap REF_L| \sim |REF_L|$.

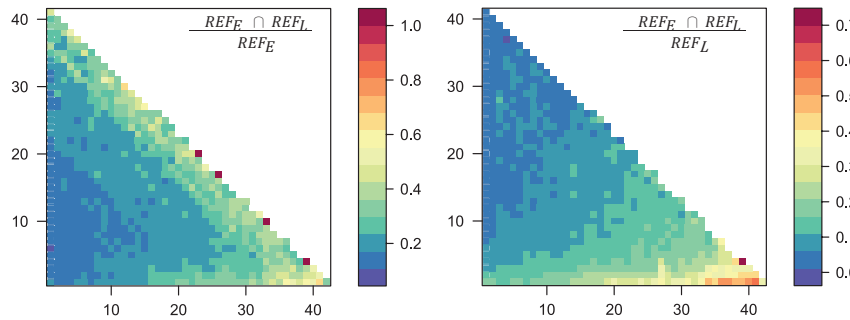


Figure 6. Average relevances of P_E (left) and P_L (right) for pairs with BCS > 1 according to ages (x) and spans (y).

Figure 7 shows the distribution of the average CR for BC pairs whose BCS is greater than 1. As illustrated, CR, as the product of one factor favoring short-spanned pairs and the other favoring long-spanned pairs, has diminished span effect. The age effect remains but to a lesser

extent compared to what is revealed in the right diagram of Figure 6. Both phenomena are perhaps due to that P_E's factor does not strictly incline towards aged pairs or long-spanned pairs, but more towards pairs having greater (age+span) values. This is why pairs having higher P_E's factor are more often distributed along the diagonal of the left diagram of Figure 6. This is also why the span effect is not entirely cancelled, and the age effect is lessened. As CR is more uniformly distributed over the ages and spans of BC pairs, a fixed threshold may be applied without causing significant discrimination against aged or long-spanned pairs.

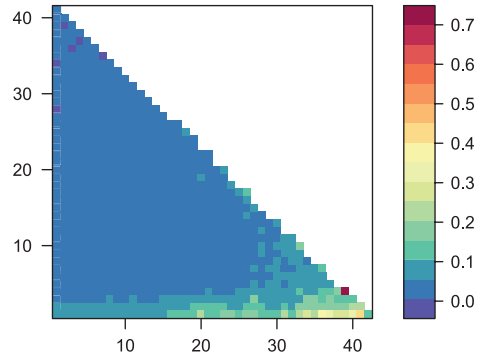


Figure 7. Average CR for pairs with BCS > 1 according to ages (x) and spans (y).

In addition to CR's relatively more uniform distribution across ages and spans, CR also retains more aged and long-spanned pairs. For all BC pairs with BCS>1, their average BCS is 7.32 with standard deviation 29.40, and their average CR is 0.043 with standard deviation 0.14. Figure 8 then provides two diagrams, one showing the frequency distributions of 49,873 BC pairs having above average BCS (left) and the other one showing 55,954 BC pairs having above average CR (right). As illustrated, BCS filters out most pairs having spans above 30, whereas a number of them are still retained by CR. BCS also has fewer larger-aged and longer-spanned pairs. The two sets of BC pairs have 27,369 pairs in common, accounting for 55% of BC pairs in the left diagram, and 49% of BC pairs in the right diagram. In other words, those in the left diagram are not a subset to those in the right diagram. For about half of the BC pairs considered to have reflected relatedness by BCS (or CR), they are indicated otherwise by CR (or BCS).

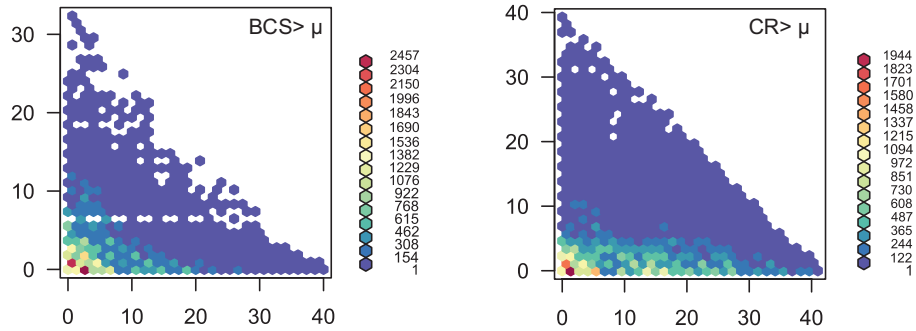


Figure 8. Frequency distributions of BC pairs having above average BCS (left) and CR (right) according to ages (x) and spans (y).

Conclusion

This study describes a temporal pattern found in bibliographically coupled patents, which indicates that BC pairs' temporal characteristics, age and span, may affect their BCS measurement, in addition to the relatedness of their patents. This study therefore proposes a new measure, combined relevance (CR), to reduce such impact.

CR is not ideal as observed above, but it is as simple as the conventional measures, both conceptually and computationally. More importantly, due to its more uniform distribution across various ages and spans, a fixed threshold may be safely applied with a reduced possibility of erroneously removing BC pairs involving truly related patents.

This study does not claim that conventional BCS measures should be replaced by CR, or CR is superior to conventional measures in every respect. When BC pairs are collected from patents issued within a time window, as their ages and spans are confined simultaneously in the same period of time, conventional measures are still viable tools. But for observing long-term knowledge dissemination or tracing overall development trajectory, CR may be a promising new alternative.

This study points out that citable patent expansion and cited reference expansion, the increasing numbers of patents issued and references cited per patent, are the two factors contributing to the temporal pattern. The citable patent expansion is particularly applicable to U.S. patents, as U.S. requires full and obligatory disclosure from patent applicants. However, there is a lack of evidence that non-U.S. patents would undergo cited reference expansion of comparable degree. Therefore, one cannot conclude confidently about non-US patents. Nonetheless, CR's division of relatedness into factors respectively reflecting the relevance of shared references to the patents is a reasonable design, and should still be applicable to non-U.S. patents as well.

This study may be improved in a number of ways. Firstly, a theoretical framework for the temporal pattern should be established, and the temporal pattern should be further verified through rigorous statistical analysis, in addition to the above general observation. Secondly, CR itself may have room for improvement. For one example, the concept of present value may be borrowed and the denominator $|REF_L|$ in $\frac{|REF_E \cap REF_L|}{|REF_L|}$ may be discounted to the date t_E to compensate the cited reference expansion. CR may also be applied to real case data, and the result is compared to that by a conventional measure to see how they differ in an application setting. Finally, the particular threshold picked for removing noise requires further investigation for its validity and influence.

Acknowledgments

This work was financially supported by the Center for Research in Econometric Theory and Applications (Grant no. 108L900204) from The Featured Areas Research Center Program within the framework of the Higher Education Sprout Project by the Ministry of Education (MOE) in Taiwan, and by the Ministry of Science and Technology (MOST), Taiwan, under Grant No. MOST 108-3017-F-002-003-.

References

Arts, S., Cassiman, B., & Gomez, J. C. (2018). Text matching to measure patent similarity. *Strategic Management Journal*, *39*(1), 62-84.

Angue, K., Ayerbe, C., & Mitkova, L. (2014). A method using two dimensions of the patent classification for measuring the technological proximity: an application in identifying a potential R&D partner in biotechnology. *The Journal of Technology Transfer*, *39*(5), 716-747.

Bicknell, G. M. (2008). To Disclose or Not to Disclose: Duty of Candor Obligations of the United States and Foreign Patent Offices. *Chi.-Kent L. Rev.*, *83*, 425.

Chang, S. B. (2012). Using patent analysis to establish technological position: Two different strategic approaches. *Technological Forecasting and Social Change*, *79*(1), 3-15.

Chen, D. Z., Huang, M. H., Hsieh, H. C., & Lin, C. P. (2011). Identifying missing relevant patent citation links by using bibliographic coupling in LED illuminating technology. *Journal of Informetrics*, *5*(3), 400-412.

Choi, S. S., Cha, S. H., & Tappert, C. C. (2010). A survey of binary similarity and distance measures. *Journal of Systemics, Cybernetics and Informatics*, *8*(1), 43-48.

Glänzel, W., & Czerwon, H. J. (1996). A new methodological approach to bibliographic coupling and its application to the national, regional and institutional level. *Scientometrics*, *37*(2), 195-221.

Hall, B. H., Jaffe, A. B., & Trajtenberg, M. (2001). *The NBER patent citation data file: Lessons, insights and methodological tools* (No. w8498). National Bureau of Economic Research.

Jaccard, P. (1901). Etude comparative de la distribution florale dans une portion des Alpes et du Jura. *Bulletin de la Société vaudoise des Sciences Naturelles*, 37, 547-579.

Jaffe, A. B. (1986). Technological opportunity and spillovers of R&D: evidence from firms' patents, profits and market value.

Jarneving, B. (2007a). Bibliographic coupling and its application to research-front and other core documents. *Journal of Informetrics*, *1*(4), 287-307.

Jarneving, B. (2007b). Complete graphs and bibliographic coupling: A test of the applicability of bibliographic coupling for the identification of cognitive cores on the field level. *Journal of Informetrics*, *1*(4), 338-356.

Kessler, M. M. (1963). Bibliographic coupling between scientific papers. *Journal of the Association for Information Science and Technology*, *14*(1), 10-25.

Kuan, C. H., Huang, M. H., & Chen, D. Z. (2018). Missing links: Timing characteristics and their implications for capturing contemporaneous technological developments. *Journal of Informetrics*, *12*(1), 259-270.

Kuusi, O., & Meyer, M. (2007). Anticipating technological breakthroughs: Using bibliographic coupling to explore the nanotubes paradigm. *Scientometrics*, *70*(3), 759-777.

Leydesdorff, L., Kushnir, D., & Rafols, I. (2014). Interactive overlay maps for US patent (USPTO) data based on International Patent Classification (IPC). *Scientometrics*, *98*(3), 1583-1599.

Lo, S. C. (2007). Patent coupling analysis of primary organizations in genetic engineering research. *Scientometrics*, *74*(1), 143-151.

Moehrle, M. G., & Gerken, J. M. (2012). Measuring textual patent similarity on the basis of combined concepts: design decisions and their consequences. *Scientometrics*, *91*(3), 805-826.

Nakamura, H., Suzuki, S., Sakata, I., & Kajikawa, Y. (2015). Knowledge combination modeling: The measurement of knowledge similarity between different technological domains. *Technological Forecasting and Social Change*, *94*, 187-201.

Niemann, H., Moehrle, M. G., & Frischkorn, J. (2017). Use of a new patent text-mining and visualization method for identifying patenting patterns over time: Concept, method and test application. *Technological Forecasting and Social Change*, *115*, 210-220.

Ortiz-de-Urbina-Criado, M., Nájera-Sánchez, J. J., & Mora-Valentín, E. M. (2018). A Research Agenda on Open Innovation and Entrepreneurship: A Co-Word Analysis. *Administrative Sciences*, *8*(3), 34.

Park, I., Jeong, Y., Yoon, B., & Mortara, L. (2015). Exploring potential R&D collaboration partners through patent analysis based on bibliographic coupling and latent semantic analysis. *Technology Analysis & Strategic Management*, *27*(7), 759-781.

Petruzzelli, A. (2011). The impact of technological relatedness, prior ties, and geographical distance on university-industry collaborations: A joint-patent analysis. *Technovation*, *31*, 309-319.

Small, H. (1973). Co-citation in the scientific literature: A new measure of the relationship between two documents. *Journal of the American Society for information Science*, *24*(4), 265-269.

Swanson, D. (1971). Some unexplained aspects of the cranfield tests of indexing performance factors. *The Library Quarterly*, *41*(3), 223-228.
Trajtenberg, M. (1990). A penny for your quotes: patent citations and the value of innovations. *The Rand Journal of Economics*, 172-187.
Von Wartburg, I., Teichert, T., & Rost, K. (2005). Inventive progress measured by multi-stage patent citation analysis. *Research Policy*, *34*(10), 1591-1607.
Wang, A. P., Hou, C. E., & Hung, S. W. (2018). Exploration of the evolution of nanotechnology from a patent co-classification perspective. *Nanotechnology Reviews*, *7*(3), 233-245.
Yeh, H. Y., Sung, Y. S., Yang, H. W., Tsai, W. C., & Chen, D. Z. (2013). The bibliographic coupling approach to filter the cited and uncited patent citations: a case of electric vehicle technology. *Scientometrics*, *94*(1), 75-93.
Yoon, B., & Magee, C. L. (2018). Exploring technology opportunities by visualizing patent information based on generative topographic mapping and link prediction. *Technological Forecasting and Social Change*, *132*, 105-117.
Zhang, L., Huang, Y., & Chen, L. (2018). Patent Citation Inflation: The Phenomena, Measurement Methods, and Proposed Solutions Based on Relative Citation Indicators. In *23rd International Conference on Science and Technology Indicators (STI 2018), September 12-14, 2018, Leiden, The Netherlands*. Centre for Science and Technology Studies (CWTS).

Studying the Scientific Mobility and International Collaboration Funded by the China Scholarship Council

Zhichao Fang[1], Wout Lamers[1], and Rodrigo Costas[1,2]

[1] *z.fang@cwts.leidenuniv.nl, w.s.lamers@cwts.leidenuniv.nl, rcostas@cwts.leidenuniv.nl*
Centre for Science and Technology Studies (CWTS), Leiden University, Wassenaarseweg 62A, Leiden, 2333AL (The Netherlands)

[2] DST-NRF Centre of Excellence in Scientometrics and Science, Technology and Innovation Policy, Stellenbosch University (South Africa)

Abstract
Every year many scholars are funded by the China Scholarship Council (CSC). The CSC is a funding agency established by the Chinese government with the main initiative of training Chinese scholars to conduct research abroad and to promote international collaboration. In this study, we identified these CSC-funded scholars sponsored by the China Scholarship Council based on the acknowledgments text indexed by the Web of Science. Bibliometric data of their publications were collected to track their scientific mobility in different fields, and to evaluate the performance of the CSC scholarship in promoting international collaboration by sponsoring the mobility of scholars. Papers funded by the China Scholarship Council are mainly from the fields of natural sciences and engineering sciences. There are few CSC-funded papers in the field of social sciences and humanities. CSC-funded scholars from mainland China have the United States, Australia, Canada, and some European countries, such as Germany, the UK, and the Netherlands, as their preferential mobility destinations across all fields of science. CSC-funded scholars published most of their papers with international collaboration during the mobility period, with a decrease in the share of international collaboration after the support of the scholarship.

Keywords
The China Scholarship Council (CSC); Scientific mobility; International collaboration; Funding acknowledgement

Introduction

The China Scholarship Council (CSC), a non-profit funding organization entrusted by the Ministry of Education of the People's Republic of China, was established in 1996[1] for the award, enrolment and administration of a series of Chinese Government Scholarship programs. These scholarship programs were set up by the Chinese government to encourage and sponsor Chinese students and scholars to study and conduct research abroad, as well as to sponsor international students and scholars to study and do research in Chinese universities[2]. According to the *provisions* issued by the Ministry of Education of China in 2007[3], the China Scholarship Council is in charge of selecting and dispatching students and scholars based on their own applications and organized experts review, and the CSC scholarship awardees are obligated to move to their targeted visiting countries and stay there to study or conduct research within the prescribed time. Moreover, for sponsored Chinese students and scholars, they are required to return to China and work for at least two years following completion of sponsorship by principle. Therefore, it is a kind of circular transnational scientific mobility (Jöns, 2007) supported by the government with specific initiatives: to train talents and promote international collaboration in some important fields, and lastly, to implement the so-called *return brain drain* (Jonkers & Tijssen, 2008). As Cao (2008) suggested, although a small but growing return migration of Chinese researchers has been seen, the whole return rate is low and many highly qualified academics still stay abroad for multiple reasons.

Therefore, the CSC scholarship is expected to play an important role in training scholars with international research experience and attracting them back to China.

Scientific mobility has been related to higher scientific impact of papers and scientists themselves (Wagner & Jonkers, 2017; Sugimoto et al., 2017; Halevi, Moed, & Bar-Ilan, 2015). Cañibano (2017) conceived it as a mechanism for the allocation of human resources in research labour market, through which *brain gain*, *brain drain*, and *brain circulation* are happening and global inequality is therefore increased (Scott, 2015). However, different from unprompted brain drain, as Cañibano & Woolley (2015) pointed out, there are numerous developing countries formulated grant policies with *train* and *attract back* rationales. The China Scholarship Council is a typical example of national funding organisations implementing this train and attack back policy, together with some countries in Latin America, such as Peru[4], which are suffering increasing brain drain too (Adams Jr, 2003). Those scholarship awardees were expected to gain international experience (Ackers, 2008) in leading countries in their subject fields, and expand transnational collaboration networks through mobility (Meyer, Kaplan, & Charum, 2001; Cañibano, Fox, & Otamendi, 2015). Jonkers and Tijssen (2008) found that the overseas experience of Chinese plant molecular life researchers who returned to their home country does have a distinct positive impact on the publication productivity; moreover, it was observed a positive correlation between researchers' overseas experience and the quantity of their corresponding transnational co-publications. Based on survey data, Scellato, Franzoni, & Stephan (2015) concluded that migrants and returnees hold larger international research networks compared to those native researchers lacking an international background. The positive effect of international mobility among countries with rich research environments on qualified international collaboration was also observed by Kato & Ando (2016).

Identifying the mobility of scientists is the first step to quantitatively analyse and evaluate the behaviours and the following impact of mobile scientists. Bibliometric data, especially the affiliation data of authors, have been widely used in identifying and tracing the trajectory of individual authors since Laudel (2003). Through this bibliometric approach, the relationship between researcher mobility and other bibliometric indicators, such as their collaborative networks, paper productivity and citations, were discussed in previous studies (Furukawa, Shirakawa, & Okuwada, 2011; Aksnes et al., 2013). In addition, on the basis of the development and application of large-scale author name disambiguation algorithm (Caron & van Eck, 2014), studies on scientific mobility became more extensive and worldwide. For instance, Scopus author-affiliation linking and author profiling were used by Moed, Aisati & Plume (2013) and Moed & Halevi (2014) to track international migration. Based on the Web of Science data, Robinson-Garcia et al. (2019) presented a taxonomy consisting of four mobility types of scientists: not mobile researchers, directional travellers, non-directional travellers, and migrants, by using instances of changes in (or multiplicity of) affiliations for a single scholar as the proxy for identifying mobility. The same method was applied by Sugimoto et al. (2017) and Chinchilla-Rodríguez et al. (2018) in tracking international mobility, under the background of controversial travel bans issued by the US government which led to a negative influence on the scientific activities of scholars from those restricted countries (Reardon, 2017; Morello & Reardon, 2017). However, there are some limitations to this methodology based on the changes of authors' affiliations, such as ignoring the scientific mobility without research outputs and underrepresenting short-term stays without changes of authors' affiliations (Robinson-Garcia et al., 2019).

The mobility and effect resulted by some international scholarship programs have been analysed based on annual report or survey data, such as the US Fulbright Program (Kahn & MacGarvie, 2011). In this paper, we study the mobility of scholars who were funded by the China Scholarship Council by mining the acknowledgements text of the Web of Science

papers. Their mobility destinations were identified through affiliation information of authorships, and all of their Web of Science publications were collected by using the author name disambiguation algorithm by Caron & van Eck (2014) to investigate the proportion of their papers with international collaboration in different periods (before, during, and after sponsorship of the CSC scholarship). The scientific mobility of Chinese CSC-funded scholars is assumed to reflect the initiatives of government for training elite academics with international experience and establishing closer international collaboration. The performance of this funding policy has significant implications for the funder agency and policy makers. The main objective of this study is to investigate the situation and performance of the China Scholarship Council in promoting brain exchange and international collaboration. Here we addressed the following research questions:

- Firstly, how is the distribution of CSC-funded papers in different subject fields? Which fields have the most papers funded by the China Scholarship Council?
- Secondly, based on the affiliation information of their research outputs, where did identified CSC-funded scholars choose to get training and establish collaboration in different fields?
- Lastly, to what extent the CSC-funded scholars engaged in international collaboration before, during, and after being supported by the China Scholarship Council?

Data and methods

Papers funded by the China Scholarship Council

According to the *Agreement on Funding the Study Abroad* that scholars have to sign with the China Scholarship Council when they are awarded the scholarship supporting them to visit or study abroad[5], they are required to acknowledge the funding from the China Scholarship Council in their research outputs conducted during the period of sponsorship. Therefore, on the basis of the published funding acknowledgements which have been indexed by the Web of Science (WoS) since 2008 for science and medicine papers, and since 2015 for social science papers (Paul-Hus, Desrochers, Costas, 2016), respectively, WoS papers funded by the China Scholarship Council can be identified.

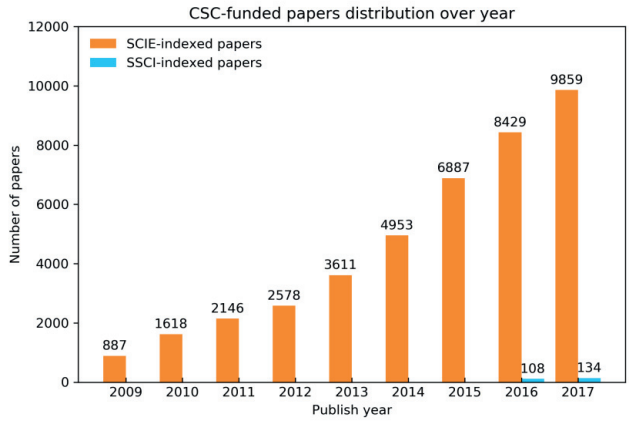


Figure 1. Temporal distribution of the CSC-funded papers.

By using the in-house version of the Web of Science maintained at CWTS of Leiden University, we collected 40,968 SCIE-indexed papers published from 2009 to 2017 and 242 SSCI-indexed papers published from 2016 to 2017 with the China Scholarship Council or its variations (such as 'the CSC Scholarship' and 'the Chinese Scholarship Council') listed in their funding sources and standardized in the 'CWTS Thesaurus' (Van Honk, Calero-Medina, & Costas, 2016). Only the document type of Article and Review are considered. The temporal distribution of these 41,210 papers is shown in Figure 1. From 2009 onwards, the number of papers funded by the China Scholarship Council increased over time. In the period from 2016 to 2017, when both SCIE-indexed papers and SSCI-indexed papers have recorded funding sources, the quantity of SCIE-indexed papers acknowledging the China Scholarship Council is much higher than SSCI-indexed papers.

Author disambiguation and identification

In this study 41,210 CSC-funded papers are contributed by 228,365 authors in total. In order to study the mobility and performance of scholars who were actually sponsored by the China Scholarship Council, firstly it is necessary to disambiguate author names to solve the problems caused by homonymy and name variants (Costas, van Leeuwen, & Bordons, 2010). A large-scale author name disambiguation algorithm developed by Caron and van Eck (2014), which has been implemented in the in-house CWTS version of the Web of Science, was used to identify unique scholars authoring scientific papers. Based on the disambiguation algorithm, we found that 41,210 CSC-funded papers were contributed by 141,614 unique individuals. However, not all of them were funded by the China Scholarship Council and supported by the Chinese government to move. Actually, most contributed individuals are co-authors of CSC-funded scholars. Therefore, in order to explore the scientific mobility and international collaboration of CSC-funded scholars, we identified those real CSC Scholarship awardees among all authors through mining the acknowledgements text, by matching the authors' names in the acknowledgements. A brief description of this text mining methodology is as follows:

- Firstly, specific full sentences containing "China Scholarship Council" or its variants, such as "CSC scholarship", "Chinese Scholarship Council", are extracted from the acknowledgement text of each CSC-funded paper. Thus, an example of an extracted full sentence is "Long Chen is supported by a scholarship from the China Scholarship Council (CSC)." (extracted from the acknowledgements of WoS paper: 000373106500002).
- Secondly, for every CSC-funded paper, various forms of its authors' names are developed (e.g. full name and last name, initials and surnames, etc.). Every author is also assigned with an ordinal number based on their sequences, such as "the first author", "the second author". Taking the WoS paper (000373106500002) as an example, there are two authors: "Bernreuther, Werner" and "Chen, Long". Possible forms of each author's name in the acknowledgements are developed as: Bernreuther, W ("Bernreuther Werner", "Werner Bernreuther", "Bernreuther. W", "B.W.", "BW", "the first author", etc.), Chen, L ("Chen Long", "Long Chen", "Chen. L", "C.L.", "CL", "the second author", etc.).
- Lastly, for each paper, the developed authors' names, together with their ordinal numbers, are matched with the corresponding extracted full funding sentence to find out if they appear in it or not. Once an author's name or ordinal number is matched, that author would be identified as the CSC-funded scholar of that paper. In the above example, the name of "Long Chen" was matched in the full funding sentence, so Chen, L was identified as the CSC-funded scholar.

Finally, among 141,614 unique authors, 9,562 of them were identified as CSC-funded scholars, contributing to 16,037 unique papers (i.e. for around 39% of the CSC-funded publications we identified the funded scholar).

Papers contributed by CSC-funded scholars

For the 9,562 identified CSC-funded scholars, all of their published WoS papers until March, 2018 were collected based on the disambiguation algorithm and further cleaned by authors' first names, affiliations, and co-authorship. In addition to the 16,037 CSC-funded papers with identified scholars, there are other 69,708 WoS papers that were authored by these scholars, thus totalling 85,745 (i.e. 16,037 CSC-funded publications with identified scholars and their other 69,708 publications that were not sponsored by the CSC scholarship or did not mention their names in the acknowledgements text).

By using the created dates of DOIs collected from Crossref as the proxy for the precise publication dates of papers (when the created date was not available (account for 9.6%), the WoS publication year was used as the alternative). As a result the 85,745 papers were classified into three periods: *before sponsorship*, *during sponsorship*, and *after sponsorship*:

- For papers with funding acknowledgements containing the China Scholarship Council, they are classified as *during sponsorship* (19,328 papers, account for 22.5%). Among these 19,328 papers, 16,037 of them mentioned the specific identified authors' names in the acknowledgements texts, while others only listed the CSC as a funding source without mentioning the CSC-funded scholars' names;
- For papers without funding acknowledgements containing the China Scholarship Council, the output of the identified CSC-funded scholars is analysed in order to estimate the first and last publications of the scholar with a CSC funding acknowledgment. Thus, those papers with publication date earlier than the first CSC-funded papers are classified as *before sponsorship* (30,399 papers, account for 35.5%), while papers with publication date later than the last CSC-funded papers are classified as *after sponsorship* (23,935 papers, account for 27.9%).

If a paper was authored by more than one identified CSC-funded scholars, as long as its acknowledgements contain the China Scholarship Council, it would be *during sponsorship*. Otherwise, it would be identified as *before sponsorship* or *after sponsorship* only if it could be classified into a specific period in all cases of scholars. There are two types of not CSC-funded papers that were excluded from this analysis (12,083 papers in total, accounting for 14.1%) since their classification is ambiguous based on the publication date. One is a group of papers that cannot unambiguously classified into before or after sponsorship since they were contributed by more than one CSC-funded scholar with different sponsorship periods, and they can be classified into different periods based on different authors' sponsorship periods. For instance, if a paper was authored by two identified CSC-funded scholars, from the view of the first author, it should be classified as *before sponsorship*, but from the view of the second author, it is *after sponsorship*, then this paper was excluded; the other group of excluded papers is that of papers published between the first and last publication dates of identified CSC-funded papers, namely the publication date of a paper is during sponsorship period but it didn't acknowledge the China Scholarship Council. There are some possible reasons for this situation. For example, if a paper was completed before sponsorship but published during sponsorship due to the publication delay, its acknowledgements would not contain the CSC scholarship. Besides, if the main work of a paper was not conducted during sponsorship, it is not necessary to acknowledge the China Scholarship Council even though it was published during sponsorship.

Results

Field distribution of CSC-funded papers

According to the outline[6] for selecting and dispatching CSC-funded scholars launched by the China Scholarship Council, candidates from the key fields that highlighted by two Chinese government policy documents (listed in Table 1) and humanities and social sciences have the priority to be funded. Key fields of natural sciences and engineering sciences account for the majority in these two Chinese government national outlines, especially those fields play significant roles in industry development, for instance, Manufacturing, Energy, Materials, and Transportation. Biotechnology, Environment, and Agriculture, which are of great concern to population health and public security, are also highlighted by these two policy documents. Some social science related fields are underlined as key fields too, such as Financial Accountancy, Education, and Politics and Law.

Table 1. Key fields stated in two Chinese Government National Outlines.

	National Outline for Medium and Long-Term Talents Development Plan (2010-2020)	*National Outline for Medium and Long-Term Science and Technology Development Plan (2006-2020)*
Key fields	Equipment Manufacturing Information Biotechnology Advanced Materials Aviation and Astronautics Ocean Financial Accountancy International Business Ecology and Environment Protection Energy Resources Modern transportation Agricultural Science and Technology Education Politics and Law Propaganda, Ideology and Culture Disaster Prevention and Reduction	Energy Water and Mineral Resources Environment Agriculture Manufacturing Industry Transportation Industry Information and Modern Service Industry Population and Health Sciences Urbanization and Urban Development Public Security National Defence

Figure 2 shows the field distribution of 41,210 CSC-funded papers and 16,037 of them with identified CSC-funded authors. Fractional-counting was applied to calculate the distribution of papers across fields when the paper belongs to more than one subject field. The classification of fields and field weights of each paper are based on the NOWT classification system (Tijssen, Hollanders, & van Steen, 2010) developed by CWTS. For most subject fields, nearly 40% of CSC-funded papers have authors identified as CSC-funded scholars according to the detailed statement in the funding acknowledgements. Physics and Materials Science is the field with most CSC-funded papers, followed by Chemistry and Chemical Engineering. Nearly 40% of total CSC-funded papers belong to these two fields, and the quantity is around twice larger than the third most productive field: Basic Life Sciences. Most key fields presented in Table 1 have a considerable number of CSC-funded research outputs, except for social sciences fields, although some social sciences are regarded as the key fields by the outline.

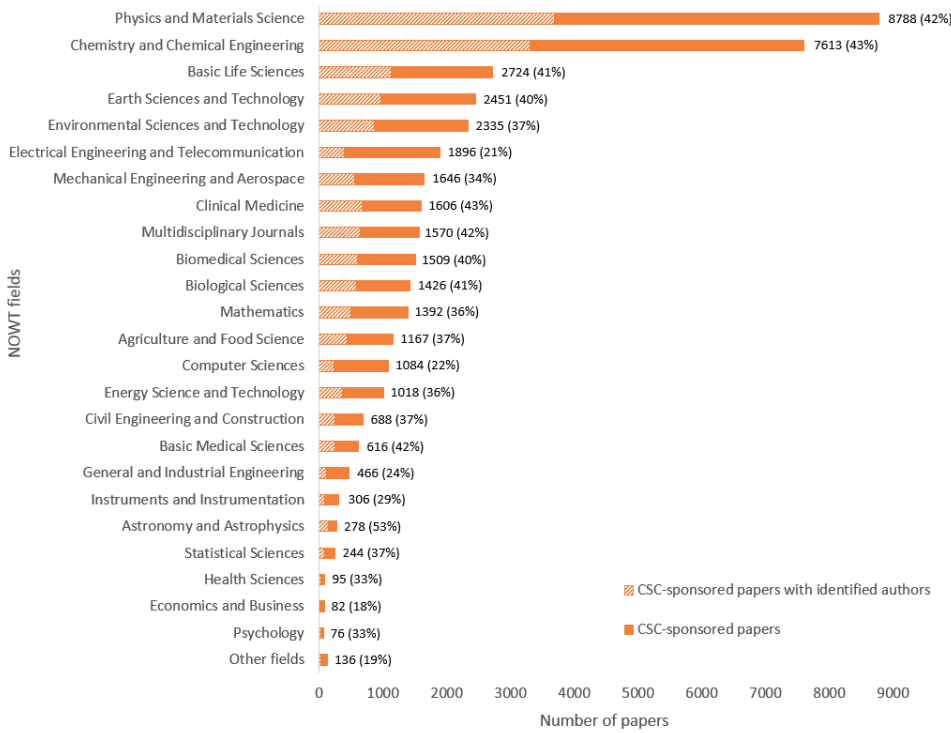


Figure 2. Field distribution of CSC-funded papers.

Scientific mobility of Chinese CSC-funded scholars

In this section we study the countries or regions where CSC-funded scholars applied and admitted to receive academic training or conduct research. Their mobility destinations not only reflect the sponsored scholars' intentions to move, but also reveal the recognition from expert reviewers organized by the China Scholarship Council to the research quality of that country in specific fields. In this study we considered both identified CSC-funded authors' affiliations and their co-authors' affiliations to identify their mobility destinations. Only identified authors from mainland China were taken into consideration to investigate their mobility destinations under the support of the China Scholarship Council, because for CSC-funded authors from other countries, their mobility destination must be China in this case. We identified Chinese CSC-funded authors by matching if they have a typical Chinese last name from mainland China (9,467 Chinese individuals were extracted, accounting for 99% of all identified CSC-funded scholars). The methodology for identifying mobility destinations are following:

- If the CSC-funded author has only one affiliated country and it is not China, then this country is the mobility destination (3,052 authors, account for 32.2%);
- If the CSC-funded author has two affiliated countries and one of them is China, then the other one is the mobility destination (4,515 authors, account for 47.7%);
- If the CSC-funded author is only affiliated to China or the author does not have affiliation information, but the co-authors are affiliated to just one another country, than that country is the mobility destination (1,233 authors, account for 12.9%);

However, there are some cases that the mobility destination cannot be identified:

- If the CSC-funded author was affiliated to more than one country except China (147 authors, account for 1.6%);
- If the CSC-funded author is only affiliated to China, while the co-authors are only affiliated to China too or affiliated to more than one country except China (814 authors, account for 8.6%).

For 8,800 Chinese CSC-funded scholars with publications whose mobility destinations can be identified more accurately as depicted above, their mobility routes are presented in Figure 3. The United Sates is the main mobility destination, together with some developed European countries, such as Germany, the UK, France, and the Netherlands. Scholars also prefer to move to Canada, Australia, and Japan to conduct research and obtain international experience.

Figure 3. Scientific mobility destinations of CSC-funded scholars from mainland China.

Table 2. Top 10 mobility destinations across five subject fields.

Rank	*Physical Sciences and Engineering*	*Biomedical and Health Sciences*	*Life and Earth Sciences*	*Mathematics and Computer Sciences*	*Social Sciences and Humanities*
1	US	US	US	US	US
2	Germany	Germany	Germany	UK	Netherlands
3	UK	Canada	Australia	Canada	UK
4	Australia	UK	Canada	Australia	Australia
5	France	Australia	UK	France	Germany
6	Canada	Netherlands	Netherlands	Germany	Canada
7	Japan	France	France	Netherlands	Spain
8	Sweden	Sweden	Japan	Japan	France
9	Netherlands	Japan	Belgium	Singapore	Denmark
10	Belgium	Belgium	Sweden	Spain	Singapore

The top 10 mobility destinations across five main disciplines (based on the CWTS classification system developed by Waltman & Van Eck, (2012)) are listed in Table 2. The field that a scholar belongs to was based on the publications during the sponsorship. Fractional counting was employed when their publications clustered into various fields. In all fields, the US is always the most preferential destination for CSC-funded scholars with

research outputs. Followed by Germany, another main choice in the fields of natural sciences and engineering sciences. Canada, Australia, and other European countries, such as the UK, the Netherlands, France, Sweden, and Belgium, also serve as important destinations of scientific mobility funded by the China Scholarship Council. Japan is the most preferential Asian country in most fields. Singapore, another Asian country, ranks in the top 10 in the field of Mathematics and Computer Sciences, and Social Sciences and Humanities.

The performance of the CSC scholarship in promoting international collaboration

Promoting international collaboration is one of the most important goals of the China Scholarship Council for sponsoring scientific mobility. We calculated the indicator *PP(IC)*, namely the proportion of papers with international collaboration (papers with affiliations from more than one country), to measure the transnational collaboration situation of papers contributed by identified CSC-funded scholars before, during, and after the sponsorship period. Figure 4 presents the proportion of papers with international collaboration across fields in different sponsorship periods.

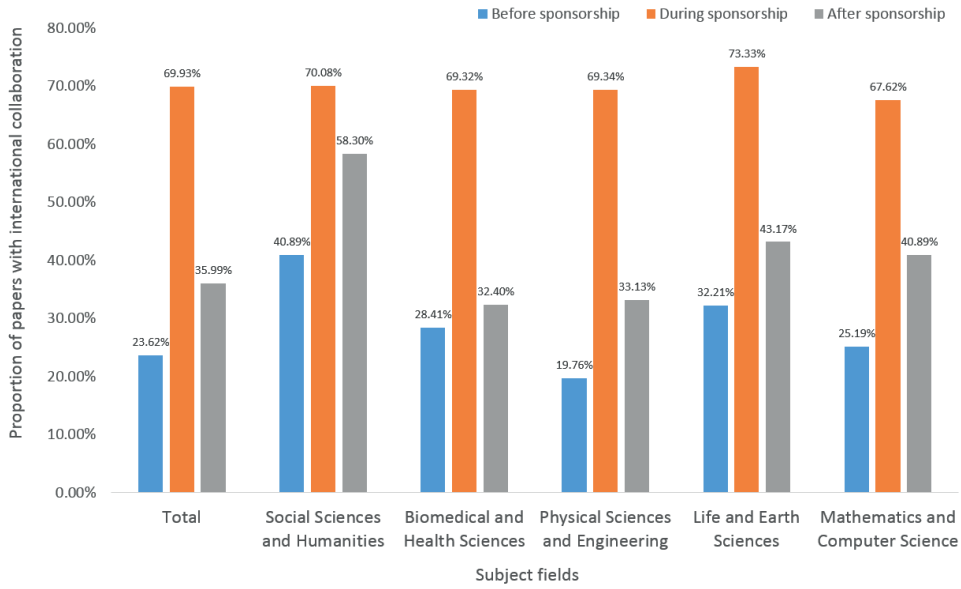


Figure 4. Proportion of papers with international collaboration in different sponsorship periods.

A similar pattern can be observed in total and across five main fields: papers published before sponsorship have the lowest rate of international collaboration, while papers published during sponsorship hold the highest proportion of international collaboration. After sponsorship, the proportion goes down compared to during sponsorship, but it is still higher than before sponsorship. These results suggest that CSC-funded scholars are most likely to collaborate with researchers from other countries during their mobility sponsored by the China Scholarship Council, and keep to some extent a relatively higher international collaboration after sponsorship. The retention rate of international collaboration after sponsorship varies across fields. For example, in the field of Social Sciences and Humanities, the proportion of papers with international collaboration during and after sponsorship are 70.1% and 58.3% respectively. In the field of Biomedical and Health Sciences, the proportion of papers with

international collaboration after sponsorship decreased sharply, from 69.3% during sponsorship to 32.4%, although still being higher than the *before* period. The same downward trend can be found in other natural sciences and engineering sciences fields. One potential reason for the high proportion of international collaboration after sponsorship might be that most collaboration relationships established during mobility were continued, suggesting that the sponsorship from the China Scholarship Council contribute to improve the international collaboration from the perspective of research outputs, especially during the period when CSC-funded scholars conducted research abroad.

Discussion and conclusions

In this study, papers with funding from the China Scholarship Council were extracted from the Web of Science to investigate the field distribution of CSC-funded papers, the scientific mobility destinations of CSC-funded scholars, and the performance of the CSC scholarship in promoting international collaboration by providing awardees with the opportunity to conduct research abroad.

For the first research question about the productivity of papers contributed by CSC-funded scholars across fields, the results of the analysis presented in this paper shows that the number of papers funded by the China Scholarship Council increased over time. The CSC-funded papers are mainly from the fields of natural sciences and engineering sciences, especially physics and material science, chemistry and chemical engineering, which were emphasized by the China Scholarship Council in the outline for selecting awardees. However, social sciences and humanities, the key field that was highlighted by the China Scholarship Council in selecting scholarship awardees as well, has much fewer WoS-covered CSC-funded papers. There are several reasons for the lower production of CSC-funded papers in social sciences. In addition to the lower coverage and shorter acknowledgement index period in the Web of Science, another possible reason is that the number of CSC-funded scholars from the fields of natural sciences and engineering sciences is larger than that from the social sciences and humanities; thus the larger number of sponsored scholars from natural sciences would also explain the larger CSC-output in these fields.

Regarding the mobility destinations of CSC-funded scholars, through text-mining the acknowledgements of CSC-funded papers, a total of 9,467 scholars from mainland China sponsored by the CSC were identified. Thus, their mobility destinations can be tracked based on their affiliation information. On the basis of affiliation information of those identified Chinese CSC-funded scholars and their co-authors, most Chinese CSC-funded scholars chose to conduct research in the USA, Australia, Canada, and other developed European countries, such as Germany, the UK, and the Netherlands, and establish collaboration relationship with researchers from these countries.

Finally, for the third research question about the performance of the China Scholarship Council in promoting international collaboration, papers contributed by CSC-funded scholars during the sponsorship period show the highest proportion of international collaboration in whichever fields. During this period, scholars have moved abroad to conduct research, it is easier for them to interact and communicate with researchers in their mobility destinations and then expand their collaboration networks. Although the rate obviously declines for papers published after sponsorship, it is still higher than papers before sponsorship. Therefore, the support from the China Scholarship Council promoted the possibility for scholars to participate in international networks. This effect partly continued when the sponsorship had ended, since papers from CSC-funded scholars still showed a relatively higher proportion of international collaboration after the sponsorship.

There are several limitations to this study. Firstly, the same as previous studies focusing on scientific mobility by using bibliometric data, only the CSC-funded scholars with research

outputs acknowledging the China Scholarship Council were identified and analysed, while those who never published WoS papers or didn't mention their funding sources are not considered, due to the lack of effective means of identifying these scholars and their performance based on bibliometric meta data. Secondly, for those papers not clearly mentioning the author who was sponsored by the China Scholarship Council in the acknowledgements, we cannot study which author is funded by the CSC scholarship. Thirdly, there might exist errors in the classification of sponsorship periods because of the publication delay. If a paper published before sponsorship was delayed for a long time, its publication date might be later than the last publication date of CSC-funded paper and the paper could be erroneously identified as *after sponsorship*. Lastly, as we mentioned above, the author name disambiguation algorithm has shortcomings in clustering Chinese names, although we have cleaned the authors' clusters data based on their Chinese first names and affiliations further, it is possible that there still exists inaccuracies.

Acknowledgments

This research is partially funded by the South African DST-NRF Centre of Excellence in Scientometrics and Science, Technology and Innovation Policy (SciSTIP). Zhichao Fang acknowledges the financial support from the China Scholarship Council (Grant No. 201706060201).

References

Ackers, L. (2008). Internationalisation, mobility and metrics: A new form of indirect discrimination?. *Minerva*, 46(4), 411-435.

Adams Jr, R. H. (2003). International migration, remittances, and the brain drain: A study of 24 labor-exporting countries. World Bank Policy Research Working Paper No. 3069. Available at SSRN: https://ssrn.com/abstract=636431

Aksnes, D. W., Rørstad, K., Piro, F. N., & Sivertsen, G. (2013). Are mobile researchers more productive and cited than non-mobile researchers? A large-scale study of Norwegian scientists. *Research Evaluation*, 22(4), 215-223.

Cañibano, C. (2017). Scientific mobility and economic assumptions: From the allocation of scientists to the socioeconomics of network transformation. *Science as Culture*, 26(4), 505-519.

Cañibano, C., & Woolley, R. (2015). Towards a socio-economics of the brain drain and distributed human capital. *International Migration*, 53(1), 115-130.

Cañibano, C., Fox, M. F., & Otamendi, F. J. (2015). Gender and patterns of temporary mobility among researchers. *Science and Public Policy*, 43(3), 320-331.

Cao, C. (2008). China's brain drain at the high end: why government policies have failed to attract first-rate academics to return. *Asian Population Studies*, 4(3), 331-345.

Caron, E., & van Eck, N. J. (2014). Large scale author name disambiguation using rule-based scoring and clustering. *In Proceedings of the 19th International Conference on Science and Technology Indicators* (pp. 79-86). Leiden: CWTS-Leiden University.

Costas, R., Van Leeuwen, T. N., & Bordons, M. (2010). A bibliometric classificatory approach for the study and assessment of research performance at the individual level: The effects of age on productivity and impact. *Journal of the American Society for Information Science and Technology*, 61(8), 1564-1581.

Chinchilla-Rodríguez, Z., Bu, Y., Robinson-García, N., Costas, R., & Sugimoto, C. R. (2018). Travel bans and scientific mobility: utility of asymmetry and affinity indexes to inform science policy. *Scientometrics*, 116(1), 569-590.

Furukawa, T., Shirakawa, N., & Okuwada, K. (2011). Quantitative analysis of collaborative and mobility networks. *Scientometrics*, 87(3), 451-466.

Halevi, G., Moed, H. F., & Bar-Ilan, J. (2015). Researchers' mobility, productivity and impact: Case of top producing authors in seven disciplines. *Publishing Research Quarterly*, 32(1), 22-37.

Jonkers, K., & Tijssen, R. (2008). Chinese researchers returning home: Impacts of international mobility on research collaboration and scientific productivity. *Scientometrics*, 77(2), 309-333.

Jöns, H. (2007). Transnational mobility and the spaces of knowledge production: A comparison of different academic fields. *Social Geography*, 2, 97-114.
Kahn, S., & MacGarvie, M. (2011). The effects of the foreign Fulbright program on knowledge creation in science and engineering. The rate and direction of inventive activity revisited (pp. 161-197). University of Chicago Press.
Kato, M., & Ando, A. (2016). National ties of international scientific collaboration and researcher mobility found in Nature and Science. *Scientometrics*, 110(2), 673-694.
Laudel, G. (2003). Studying the brain drain: Can bibliometric methods help?. *Scientometrics*, 57(2), 215-237.
Meyer, J. B., Kaplan, D., & Charum, J. (2001). Scientific nomadism and the new geopolitics of knowledge. *International Social Science Journal*, 53(168), 309-321.
Moed, H. F., & Halevi, G. (2014). A bibliometric approach to tracking international scientific migration. *Scientometrics*, 101(3), 1987-2001.
Moed, H. F., Aisati, M., & Plume, A. (2013). Studying scientific migration in Scopus. *Scientometrics*, 94(3), 929-942.
Morello, L., & Reardon, S. (2017). Meet the scientists affected by Trump's immigration ban. *Nature*, 542(7639), 13.
Paul-Hus, A., Desrochers, N., & Costas, R. (2016). Characterization, description, and considerations for the use of funding acknowledgement data in Web of Science. *Scientometrics*, 108(1), 167–182.
Reardon, S. (2017). How the latest US travel ban could affect science. *Nature*, 550(7674), 17.
Robinson-Garcia, N., Sugimoto, C. R., Murray, D., Yegros-Yegros, A., Larivière, V., & Costas, R. (2019). The many faces of mobility: Using bibliometric data to measure the movement of scientists. *Journal of Informetrics*, 13(1), 50-63.
Scellato, G., Franzoni, C., & Stephan, P. (2015). Migrant scientists and international networks. *Research Policy*, 44(1), 108-120.
Scott, P. (2015). Dynamics of academic mobility: Hegemonic internationalisation or fluid globalisation. *European Review*, 23(S1), S55-S69.
Sugimoto, C. R., Robinson-García, N., Murray, D. S., Yegros-Yegros, A., Costas, R., & Larivière, V. (2017). Scientists have most impact when they're free to move. *Nature*, 550(7674), 29.
Tijssen, R., Hollanders, H., van Steen, J. (2010). Wetenschaps en Technologie Indicatoren 2010. Nederlands Observatorium Wetenschap en Technologie (NOWT).
Van Honk, J., Calero-Medina, C., & Costas, R. (2016). Funding Acknowledgements in the Web of Science: inconsistencies in data collection and standardization of funding organizations. *In Proceedings of the 21st International Conference on Science and Technology Indicators*. Book of Proceedings.
Wagner, C. S., & Jonkers, K. (2017). Open countries have strong science. *Nature*, 550(7674), 32.
Waltman, L., & Van Eck, N. J. (2012). A new methodology for constructing a publication-level classification system of science. *Journal of the American Society for Information Science and Technology*, 63(12), 2378-2392.

---

[1] Introduction to China Scholarship Council (in Chinese): https://www.csc.edu.cn/about
[2] Introduction to Chinese Government Scholarships: http://www.campuschina.org/content/details3_74776.html
[3] Provisions on the Management of Postgraduates Funded by China Scholarship for Studying Abroad (for Trial Implementation) (in Chinese): http://www.moe.edu.cn/s78/A20/s7068/201410/t20141021_178464.html
[4] Científicos lamentan estado de la ciencia en Perú (in Spanish): https://www.scidev.net/america-latina/migracion/noticias/cient-ficos-lamentan-estado-de-la-ciencia-en-per-.html
[5] Provisions on Several Issues Concerning the Dispatch and Management of Personnel Funded by China Scholarship for Studying Abroad (1996) (in Chinese): https://www.csc.edu.cn/chuguo/s/168
[6] 2018 Outline of selecting and dispatching CSC-funded scholars (in Chinese): https://www.csc.edu.cn/article/1042

Institutional research specializations identified by esteem factors and bibliometric means: A case study at the University of Vienna

Johannes Sorz[1], Wolfgang Glänzel[2], Ursula Ulrych and Juan Gorraiz[3]

[1] johannes.sorz@univie.ac.at
University of Vienna, Office of the Rectorate, Universitätsring 1, A-1010 Vienna (Austria)

[2]wolfgang.glanzel@kuleuven.be, glanzw@iif.hu
ECOOM and Dept MSI, KU Leuven, Naamsestraat 61, Louvain, 3000 (Belgium)
Library of the Hungarian Academy of Sciences, Dept. Science Policy and Scientometrics, Arany János u. 1, Budapest, 1051 (Hungary)

[3] ursula.ulrych@univie.ac.at, juan.gorraiz@univie.ac.at,
University of Vienna, Vienna University Library, Dept of Bibliometrics, Boltzmanngasse 5, A-1090 Vienna (Austria)

Abstract

The identification of one's own research specializations is of crucial importance for research administration at universities. In this case study, two different approaches were applied to the University of Vienna. The first relies on third-party funding and ranking results as well as on other esteem indicators. The second is based on a bibliometric analysis of the publication output. We used the Web of Science subject categories for publications clustering. In addition publications were also clustered for two esteem indicator-based research specializations (one STEM, one SSH) by means of the lists of the researchers associated with the research specializations (RS).Both approaches proved to be useful in identifying research specializations, lead to similar results and are meant to be used in a complementary way. We found that the greatest hindrance lies in the inherent limitations of journals-assignment-based classification systems that does not correspond to university structure and the considerable time and efforts for more accurate researcher-based publication analyses. Further investigation on this subject, including new and alternative metrics, is needed and will be conducted in the future. However, the preliminary results already demonstrate the benefits of bibliometric analyses for the identification of research specializations of universities.

Background & Introduction

The identification of strengths and weaknesses in research and teaching is imperative for a modern international university, especially in times of increased autonomy and economic strain (Fumasoni and Huisman, 2013). There are various attempts to identify institutional research specializations documented in literature, e.g. by van Vught and Huisman (2014), who stress the importance of defining institutional research profiles for Universities and propose a multi-faceted approach combining mapping, multi-dimensional ranking, benchmarking and degree profiling to define these. To define a multi-faceted research profile for the University of Vienna is a seminal task of its leadership and the starting point of this case study. With its 6,700 academics, more than 400 of whom are professors, the University of Vienna is the largest research and educational institution in Austria. The vast array of disciplinary and interdisciplinary research at the University of Vienna includes theology, law, business, economics and statistics, computer science, historical and philological studies, social sciences, psychology, life sciences, natural sciences, mathematics, sport science and teacher education. Research and teaching are organized in 15 Faculties and 5 centres. Each of these faculties and centres defines its unique set of key-research in a bottom-up approach by its researchers during a periodical strategy process. In total, the University of Vienna has

currently 100 key-research areas. The research profile can therefore not easily be restricted to just a few areas because there also always has to be institutional responsibility towards all other academic fields. It is nevertheless necessary to focus on the question of what the University of Vienna stands for in a special way and which research areas need to be emphasised. This question is relevant for the public image, for the correct adjustment of the self-perception and for resource decisions (Laudel and Weyer, 2014, Huisman and van Vught, 2009). In 2016, the Rectorate started a project to identify and define the so-called interdisciplinary research specializations of the University. The Federal Ministry of Education, Science and Research proposed the use of esteem indicators to assess these specializations (Burkert, 2016). Based on this proposal the Rector's Office developed a specific set of esteem indicators (see Methodology section) mainly based on competitive third party funding, science awards. Despite the well-known limitations of higher education rankings (Sorz, et al. 2015; Bookstein et al. 2010; Hazelkorn 2007) results of subject rankings were also included. While this procedure found general acceptance within the University, publication output and citations, other relevant indicators for scientific accomplishment, were so far not included. There could be areas the researchers of the University excel, but we have not considered yet as a part of a research specialisation, since there is no significant third party funding (due to either lack of activity or lack of available funding). This could be especially relevant for the Social Sciences and Humanities (SSH) that are well represented at the University of Vienna. Approximately 40 % of the total full-time equivalent scientific staff at the University of Vienna can be attributed to the humanities and approximately 22% to the social sciences. This made it imperative to investigate possible gaps in coverage of the current esteem-factor approach. Furthermore, the inclusion of publication and citation data would give a better multidimensional picture of the University's research strength and could back-up the existing esteem indicators (García et. al., 2013; Rafols, et al., 2010).

Aims of the study

The main research questions of this study are:

1) Does a bibliometric analysis confirm the established esteem-factor approach used by the University of Vienna to define its research specializations?
2) What set of bibliometric indicators complements the esteem indicators?
3) Are there research areas that not covered by the esteem-factor approach, that have potential to be new research specializations and how would the use of additional bibliometric indicators redefine the current specializations?

To answer questions 1-3 we compared the esteem-factor based research specializations and their underlying key-research areas with a bibliometric analysis of high publication numbers, high publication numbers in top journals and high impact. We used two approaches to cluster the publications. First, we used the subject categories as assigned in the Web of Science Core Collection. Then we used a list of the researchers associated with the esteem factors (awardees, grantees, and the researchers in their third-party-funded projects) of two research specializations, one STEM and one SSH.

Methodology

In order to determine the research specializations of the University we used the existing 100 key research areas of the University's Faculties and combined them with output indicators (esteem factors). The key-research areas have been defined by the researchers of the faculties

in a "bottom-up" discussion process and are laid out in the University's development plan[1]. We matched the existing key-research areas with a set of esteem factors. As esteem factors we considered only running projects and the most recent ranking results:

National grants/prizes

- FWF-START-Grant: Grant by the largest Austrian Funding Agency (FWF) for outstanding young researchers of any discipline to establish a research group (800k-1.2 Mio. €).
- Wittgenstein-Award: Highest endowed research prize in Austria (up to 1.5 Mio. €)
- FWF-Special Research Programmes (SFB) for the establishment of interdisciplinary research networks (up to 1 Mio. € per year/for up to 8 years)
- FWF-Doctoral Programmes (DK) provide funding for 5-20 PhD students for up to 8 years
- WWTF-Grants: competitive grants provided by the Vienna Science and Technology Fund Projects, endowed chairs and Vienna Research Groups in various research fields (up to 1 Mio. Euro/project)
- CD-Labs: application-oriented research funding, connecting business and science; research groups (5-15 people) up to 7 years (110k-700k/year; 50-60% public funding, 40-50% private funding by companies)
- Laura-Bassi-Centres: Governmental funding for predominantly female research teams with a female head on the interface academia/industry (ca. 1 Mio./centre/year for up to 7 years)

International grants

- ERC-Grants (Starting, Consolidator and Advanced Grants; Proof of Concept Gants)
- Collaborative EU-Projects (FP7, Horizon2020) > 300 k€ (portion of the UoV)
- ESFRI/ERIC-Infrastructures
- Innovative Medicines Initiative (IMI) Projects

Higher education Rankings (TOP 200)

- THE Ranking by subject
- QS-Ranking by subject
- ARWU (Shanghai) Subject Ranking

Additional factors

- Special facilities with high national or international visibility (e.g. the European Law Institute of the UoV)
- Endowed chairs/professorships

We identified 37 key-research areas that feature at least one esteem-factor and aggregated these into nine "research specialisations". We defined these specializations by grouping thematically or methodologically similar or adjacent key-research areas and named each research specialization to provide an overarching theme. The order in which we list the research specialisations (RS 1-9) follows the principle of going from the theoretical to the concrete and, in turn, going from the smallest to the biggest scale level. Next, we tried to match the esteem-based research specializations with bibliometric indicators. For this, we used following three premises to decide if a category can be a potential research strength or not:

1) High publication activity. We measured the number of publications indexed in the Web of Science Core Collection. We retrieved all the document types indexed in WoS Core

[1] https://www.univie.ac.at/fileadmin/user_upload/startseite/Dokumente/Entwicklungsplan2025_EN.pdf

Collection between 2010 and 2017, affiliated to the University of Vienna and analysed them according to the Web of Science category and research areas. We applied a threshold for the different publication cultures with a minimum of 100 publications in the STEM fields and of 60 for the Social Sciences and Arts & Humanities for the eight years period.

2) Large number of publications in top journals according to the subject category. We measured by the percentage of documents in Q1 journals according to their impact factor that means publications in journals with an impact factor (JIF) ranged between the top 25% in at least one of the corresponding WoS categories to which this journal has been assigned according to JCR. We used the JIF of the JCR edition of the publication year for all analysed publications. This parameter can be very helpful for the more recent publications years where the citation window is still too short in comparison to the half-cited life in order to provide a significant statement (Gorraiz & Gumpenberger, 2012; Gorraiz et al. & 2017).

3) High citation impact. In order to assess the impact we used three normalized citation indicators, to account for differences in citation cultures across research fields and the differences in citation windows. Normalized citation indicators are the 'Category Normalized Citation Impact' (CNCI) and the percentage of publications in the Top 1% and Top 10% most cited publications of the corresponding WoS Category and in the same publication year (percentile scores). The CNCI provides the citation impact (citations per paper) normalized for subject, year and document type. A publication with a CNCI value of 1.20 is 20% above the expected citation rate in the corresponding WoS Category and in the same publication year, and a publication with a value of 0.80 is 20% below the world expected citation rate. For a collection of publications, the mean value of all CNCI values is calculated. However, the CNCI is again an average or mean value and does not consider the characteristic skewness of the citation distribution. The percentiles represent the citation count threshold for different percentile cuts for each field and year. For instance, the 10th percentile represents the number of the Top 10% most cited papers in the corresponding category and for the current publication year. Top 10% is considered as a measure of 'excellence'. We calculated percentile scores according to the baselines published in InCites for the corresponding WoS Categories. However, the CNCI as well as the percentage documents in the Top 10% and Top 1% provides rather complementary information (Gorraiz & Gumpenberger, 2015 & 2016, Moed, 2017).

To match the esteem-factor approach with bibliometrics we have opted for a "hard approach" including all three indicators together. Thus, we selected only categories as potential research strengths if the values of all three indicators exceed the total average value retrieved for all the publications of that institution and if they exceed their expected values: CNCI >1, percentage of Top 10% and Top 1% higher than 10% and 1% respectively. We used InCites for all bibliometric analyses.

Furthermore, we used two different approaches in order to cluster the publications. For the approach one we used the subject categories as assigned in the WoS Core Collection. For the second approach we identified the scientists' names associated with the esteem factors of two research specializations ("*Materials and the quantum level*", RS2 and "*Construction of identity and concepts of society*", RS6) and retrieved their publications in WoS Core Collection. These

were analysed in InCites in order to retrieve all the indicators for each involved WoS subject category.

Results

Using esteem factors only, we identified nine research specializations of the University of Vienna (RS1-9):

- RS 1: *Models and algorithms*
  Key-research areas: "Algorithms, Software and Computing Technologies", "Analysis, geometric structures and mathematical physics", "Arithmetic, algebra and discrete mathematics", "Biomathematics and Dynamic Systems", "Computational Sciences", "Particle Physics, Gravitational Physics and Mathematical Physics", "Stochastics und Financial Mathematics", „Logic". Esteem factors: 1 ERC, 4 FWF-START, 2 FWF-SFB, 1 FWF-DK, 1 EU-Projects, 2 IMI-Projects, 6 WWTF-Projects; 1 Berta-Karlik-Chair, 1 Chair by Simonis-Foundation; 4 Research Platforms, QS-Ranking, ARWU-Ranking

- RS 2: *Materials and the quantum level*
  Key-research areas: "Functional and sustainable materials chemistry", "Complex Nanoscale Matter", "Quantum Optics, Quantum Nanophysics and Quantum Information". Esteem factors: 4 ERC, 2 FWF-START, 2 FWF-SFB, 3 EU-Projects, 2 CD-Labs, ARWU-Ranking, QS-Ranking, THE Ranking

- RS 3: *Molecules, cells and their interaction*
  Key-research areas: „Chromosome-Dynamics", "Computational Chemistry and und Biomolecular Simulation", „Integrative Structural Biology", „RNA-Biology", "Cell-Signalling". Esteem factors: 7 ERC, 3 FWF-START, 4 FWF-DK, 5 FWF-SFB, 2 WWTF; 2 Berta-Karlik-Chairs), 1 Laura Bassi Centre, 1 CD-Lab, 3 WWTF-Projects (2014), 1 EU-Project, THE-Ranking, 2 Research Platforms

- RS 4: *Food and drugs*
  Key-research areas: "Biological and medicinal chemistry", "Drug Discovery from Nature", "Food chemistry and Physiological chemistry". Esteem factors: 3 ERC, 1 FWF-SFB, 1 FWF-DK, 1 WWTF-Project, 2 EU-Projects, 4 IMI-Projects, 3 CD-Labs, 4 Research Platforms

- RS 5: *Microbiology, ecosystems and evolution*
  Key-research areas: "Evolution of Organismal Complexity", "Microbial Ecology and Ecosystems", "Patterns and Processes in Plant Evolution and Ecology". Esteem factors: 8 ERC; 1 Wittgenstein-Award, 1 FWF-DK, 1 FWF- START, 1 EU-Project, 2 WWTF-Projects, 4 Research Platforms, THE-Ranking, Shanghai-Ranking

- RS 6: *Construction of identity and concepts of society*
  Key-research areas: "Community, Conflict, Integration", „Global Cultures and Identities", "Cultures of the Euro-Mediterranean region and antiquity studies", "Migration, Citizenship and Belonging", "Austrian politics and Europeanisation" Esteem factors: 7 ERC, 1 FWF-SFB, 1 FWF-START; 1 Wittgenstein-Award, "Austrian Centre for Digital Humanities"; 1 ESFRI-ERIC-Infrastructure, 1 WWTF-Project, 3 Research Centres, 3 Research Platforms, QS-Ranking

- RS 7: *Cognition, communication and systemic reflection*
  Key-research areas: "Justification and critique of norms in ethics, law and politics", "Cognitive Science and Neuroscience", "Systemic and functional dimensions of communication", „Technologies and cognitive processes in translation and multilingual language data processing", "Theories of knowledge, of science and of the social". Esteem factors: 6 ERC, 1 FWF-SFB, 3 FWF-DK, 1 FWF-START; 5 WWTF-Projects, 2 Berta-Karlik-Chairs, 1 ESFRI-ERICInfrastructure, 1 EU-Project, 3 Research Platforms, QS-Ranking, THE-Ranking, Shanghai-Ranking

- RS 8: *Internationalisation of the economy and law*
  Key-research areas: "Europeanisation of commercial law and business law", "Changing markets and institutions", "Fundamental rights and human rights in complex relations of power". Esteem factors: 4 FWF-DK, 1 CD-Lab, 1 EU-Project, 1 Endowed Chair "Human Rights", "The European Law Institute"; QS-Ranking, THE-Ranking; 2 Research Platforms, 1 Research Centre

- RS 9: *The environment and cosmic processes*
  Key-research areas: "Physics and the Environment", "Cosmos", "Environment"
  Esteem factors: 2 ERC, 1 FWF-NFN, 7 EU-Projects, 1 WWTF-Project, 4 Research Platforms, QS-Ranking

The results originating from the bibliometric analyses are summarised in following tables 1 to 6. Table 1 shows the correlation between the four indicators described in the section methodology for all the Web of Sciences categories in which the University of Vienna published between 2010 and 2017.

Table 1. Correlation between the four indicators according to their values in the 236 WoS subject categories where the University of Vienna published between 2010 and 2017

	Percentage Documents in Q1 Journals	Category Normalized Citation Impact	Percentage Documents in Top 10%	Percentage Documents in Top 1%
% Documents in Q1 Journals		0.13	0.43	0.25
Category Normalized Citation Impact	0.13		0.52	0.68
% Documents in Top 10%	0.43	0.52		0.32
% Documents in Top 1%	0.25	0.68	0.32	

The very low correlation values between the percentage of documents in Q1 journals and the three normalized citation counts can be explained by the fact that there are no Q1 journals in the Arts & Humanities. Table 2 provides a more detailed view of the relationship between visibility (percentage of documents in Q1 journals) and impact represented by the percentiles. We calculated correlations for the total period (2010 and 2017) as well as for the period between 2010 and 2014 in order to guarantee a larger citation window, which is of greater value in disciplines with longer half-lives of citations.

Table 2. Distribution of the percentage of documents in Q1 journals and the citation percentile for two periods: 2010-2017 and 2010-2014

PY=2010-2017

Most Cited	*Q1*	*Q2*	*Q3*	*Q4*	*NA*
Top 1%	3.68%	0.99%	0.52%	0.35%	1.02%
[Top 1% - Top 5%]	9.63%	4.04%	2.32%	1.27%	2.67%
[Top 5% - Top 10%]	9.85%	5.61%	3.44%	1.20%	3.82%
[Top 10% - Top 20%]	15.38%	11.64%	8.68%	3.80%	5.92%
[Top 20% - Top 50%]	29.66%	32.27%	25.07%	15.97%	14.93%
[Top 50% - Top 100%]	17.81%	28.01%	32.79%	37.79%	10.12%
uncited	13.99%	17.44%	27.19%	39.62%	61.52%
Total	10412	5045	2501	1421	5393
	PY=2010-2014				
Most Cited	*Q1*	*Q2*	*Q3*	*Q4*	*NA*
Top 1%	3.71%	1.07%	0.45%	0.12%	1.18%
[Top 1% - Top 5%]	9.94%	4.29%	2.27%	1.61%	2.84%
[Top 5% - Top 10%]	10.18%	6.05%	3.51%	1.61%	3.83%
[Top 10% - Top 20%]	15.78%	12.19%	7.93%	4.38%	6.74%
[Top 20% - Top 50%]	31.03%	32.44%	27.10%	17.30%	18.27%
[Top 50% - Top 100%]	18.47%	31.24%	38.60%	45.44%	13.51%
uncited	10.90%	12.71%	20.14%	29.53%	53.61%
Total	5984	3076	1539	867	3130

The percentage of documents in Q1 journals does not strongly correlate with the citation percentiles. Furthermore, almost half of the publications do not contain this quartile information because they were not published in journals or the publishing journals do not have an impact factor in JCR. Table 2 shows that the percentage of documents in Q1 journals in top percentiles increases considerably with the citation window. The median correlation between the CNCI and the percentage of documents in Top 10% reported in Table 1 underpins the simultaneous use of both indicators, due to the skewness of the citation distribution. The correlation is higher as expected between the CNCI and the percentage of documents in Top 1%, because the outliers (extremely highly cited publications) strongly shapes the CNCI.

Table 3 shows all WoS Categories with exceeding values of the three selected indicators in comparison to the mean baseline for the overall university (baseline for all items) and with a threshold of at least 100 publications. We sorted them according to the descending percentage of documents in Top 10% most cited. As additional information, we have added the research specialization (RS) - as identified in the previous approach - that would correspond thematically to the WoS category (last column).

Table 3. Top WoS categories according to the three selected indicators (CNCI, Top 10%, Top1%)

WoS Category	*Web of Science Documents*	*Percentage Documents in Top 10%*	*Category Normalized Citation Impact*	*Percentage Documents in Top 1%*	*RS*
Baseline for All Items	***24772***	***14.33***	***1.46***	***2.03***	
MICROBIOLOGY	522	27.97	2.03	5.75	RS3/5
ECOLOGY	707	25.88	1.91	5.66	RS5/9
CARDIAC & CARDIOVASCULAR SYSTEMS	117	25.64	5.48	6.84	
SURGERY	118	23.73	5.86	7.63	
METEOROLOGY & ATMOSPHERIC SCIENCES	269	23.05	2.42	7.06	RS9

BIOLOGY	333	22.22	1.8	3.6	RS3/4/5
PHYSICS, MULTIDISCIPLINARY	680	21.91	2.22	3.38	RS2/1
BIODIVERSITY CONSERVATION	142	21.83	1.9	4.93	RS5
COMMUNICATION	232	20.69	2.32	5.6	RS7
OPTICS	337	19.29	1.73	4.75	RS2
BIOTECHNOLOGY & APPLIED MICROBIOLOGY	273	19.05	1.59	3.66	RS3
GEOGRAPHY, PHYSICAL	183	18.58	1.51	3.28	RS9
ENVIRONMENTAL SCIENCES	500	18.4	1.51	2.6	RS9
INSTRUMENTS & INSTRUMENTATION	131	18.32	1.75	3.82	
BIOCHEMISTRY & MOLECULAR BIOLOGY	1250	18.24	1.68	3.36	RS3/4
EVOLUTIONARY BIOLOGY	575	18.09	1.71	2.78	RS5
CELL BIOLOGY	424	17.69	1.68	2.83	RS3
PSYCHOLOGY, EXPERIMENTAL	282	17.38	2.39	5.32	RS7
COMPUTER SCIENCE, THEORY & METHODS	457	17.07	1.6	2.19	RS1
GENETICS & HEREDITY	436	16.06	1.59	3.44	RS3
BUSINESS	177	15.82	1.69	2.26	RS8
LINGUISTICS	234	15.38	2	3.42	RS7
ANTHROPOLOGY	241	15.35	1.65	2.07	RS6
HUMANITIES, MULTIDISCIPLINARY	157	14.65	2.12	6.37	RS6/7

The analysis shows that the inclusion of the percentage of documents in Q1 journals can be used as additional selection criterion to help identify potential research strengths. Furthermore, it is in good agreement with the normalized citation indicators. Nearly all Top-WoS categories can be associated with the esteem-based RS. Thus, the results of the bibliometric analysis confirm the esteem-factor-based approach largely. However, there are some deviations. Most strikingly, the subject category of "Mathematics" is completely missing in the bibliometric analysis. While the RS1 "*Models and algorithms*" is strongly associated with the Mathematics faculty, this subject category does not show up in the bibliometric analysis. On the other hand, some of the Top-WoS categories are not covered at the University of Vienna, e.g. "Cardiac and Cardiovascular Systems" and "Surgery" (the University of Vienna has no Medical Faculty) und thus cannot be assigned to the existing RS without further analysis, e.g. identification of the researchers involved. "Physical Geography" could be associated with RS9 and "Optics" with RS2, however there exist no underlying key-research areas at the University.

Table 4. Top WoS categories according to the three selected indicators and the percentage of documents in Q1 journals for the STEM research areas

WoS Category	*Web of Science Documents*	*Percentage Documents in Q1 Journals*	*Percentage Documents in Top 10%*	*Category Normalized Citation Impact*	*Percentage Documents in Top 1%*
ANTHROPOLOGY	241	59.79	15.35	1.65	2.07
ENVIRONMENTAL SCIENCES	500	67.4	18.4	1.51	2.6
EVOLUTIONARY BIOLOGY	575	68.43	18.09	1.71	2.78
CELL BIOLOGY	424	64.32	17.69	1.68	2.83
GEOGRAPHY, PHYSICAL	183	69.51	18.58	1.51	3.28
BIOCHEMISTRY & MOLECULAR BIOL	1250	65.27	18.24	1.68	3.36
PHYSICS, MULTIDISCIPLINARY	680	75.42	21.91	2.22	3.38
GENETICS & HEREDITY	436	59.85	16.06	1.59	3.44

BIOLOGY	333	79.74	22.22	1.8	3.6
BIOTECHNOLOGY & APPLIED MICROBIOL	273	71.43	19.05	1.59	3.66
OPTICS	337	82.24	19.29	1.73	4.75
BIODIVERSITY CONSERVATION	142	68.46	21.83	1.9	4.93
ECOLOGY	707	76.67	25.88	1.91	5.66
MICROBIOLOGY	522	76.91	27.97	2.03	5.75
CARDIAC & CARDIOVASCULAR SYSTEMS	117	69.81	25.64	5.48	6.84
METEOROLOGY & ATMOSPHERIC SCI	269	62.6	23.05	2.42	7.06
SURGERY	118	80.53	23.73	5.86	7.63

We identified all involved scientists affiliated to the University of Vienna for "*Materials and the quantum level*" (RS2) and "*Construction of identity and concepts of society*" (RS6) and retrieved their publications for the period 2010-2017. Table 5 and 6 show the results for RS2 and RS6 respectively. We sorted the WoS Categories according to the number of Web of Science Documents.

Table 5. Results for the first selected research focuses RS2 ("Materials and the quantum level")

WoS Category	*Web of Science Documents*	*Percentage Documents in Top 10%*	*Category Normalized Citation Impact*	*Percentage Documents in Top 1%*	*Percentage Documents in Q1 Journals*
Baseline for All Items	*998*	*24.35*	*2.11*	*4.11*	*76.88*
PHYSICS, MULTIDISCIPLINARY	220	41.36	4.27	8.64	92.52
MATERIALS SCIENCE, MULTIDIS.	274	23.72	1.67	1.82	80.51
PHYSICS, APPLIED	229	24.02	1.69	3.06	76.70
PHYSICS, CONDENSED MATTER	210	24.29	1.57	1.90	58.13
CHEMISTRY, PHYSICAL	212	18.87	1.40	2.36	73.93
PHYSICS, ATOMIC, MOL. & CHEMICAL	144	25.00	1.62	3.47	75.74
OPTICS	115	26.96	2.05	8.70	94.37
NANOSCIENCE & NANOTECHNOLOGY	67	19.40	1.21	0.00	94.55
CHEMISTRY, MULTIDISCIPLINARY	68	10.29	0.76	0.00	89.74
POLYMER SCIENCE	60	10.00	1.14	1.67	91.53

The datasets comprises of 50 scientists and 998 publications for RS2 (STEM), and 137 scientists and 1116 publications for RS6 (SSH). For RS2 the values of all four applied indicators are at least twice as high as expected and considerably higher as for the whole University (compare baselines in Table 5 and Table 3). In contrast to this finding, the figures for RS6 are much lower. This is due to the WoS Category "Religion", where almost 550 publications were book reviews, which have no impact. Tables 5 and 6 help to identify the main categories responsible for the high impact in the identified research specialisations. For RS2 they are "Physics, Multidisciplinary" and "Optics", for RS6 "Political Science" and "Communication". In the case of "Physics, Multidisciplinary" the values for the RS2 are twice as high as for the whole university (see Table 3) and four times higher as the expected values. For RS 6 "Communication" and "Demography" and "Political Science" show high values of the three applied indicators (note: Political Science is not included in Table 3; CNCI lower than the baseline for all the faculty 1.4; 15.4% and 1.3% in the Tops 10% and 1% respectively).

Table 6. Results for the second selected research focus RS6 ("Construction of identity and concepts of society")

WoS Category	*Web of Science Documents*	*Percentage Documents in Top 10%*	*Category Normalized Citation Impact*	*Percentage Documents in Top 1%*	*Percentage Documents in Q1 Journals*
Baseline for All Items	*1116*	*10.30*	*0.91*	*1.34*	*36.27*
RELIGION	600	0.67	0.06	0.00	0.00
POLITICAL SCIENCE	180	30.00	2.36	2.78	39.64
COMMUNICATION	67	28.36	2.66	4.48	36.36
HISTORY	65	10.77	0.87	0.00	16.67
ECONOMICS	53	26.42	1.80	0.00	41.67
AREA STUDIES	37	13.51	0.83	0.00	52.38
SOCIOLOGY	34	8.82	0.66	0.00	66.67
PUBLIC ADMINISTRATION	27	11.11	0.82	0.00	50.00
DEMOGRAPHY	22	31.82	3.41	9.09	26.67
SOCIAL SCIENCES, INTERDISCIPLINARY	21	33.33	3.66	14.29	55.56
INTERNATIONAL RELATIONS	19	15.79	2.39	5.26	92.86
LANGUAGE & LINGUISTICS	18	0.00	0.15	0.00	0.00
ANTHROPOLOGY	16	25.00	1.23	0.00	16.67
Women's Studies	12	25.00	1.66	0.00	40.00
LAW	12	0.00	0.75	0.00	28.57

Discussion and conclusions

Our case study shows that bibliometric indicators offer a useful tool set for the identification of research specializations. It can therefore serve as an affirming complement of the esteem-factor approach.. We strongly advice the use of multifaceted approaches for this purpose. The combination of both approaches will most likely yield more acurate results.

Normalised citation counts are the most adequate indicators for this purpose. The percentage of documents in the "Top 10% most cited" seems to be a veryrelevant indicator. However, the CNCI and the Top 1% are helpful in addition to provide a more complete and accurate description of the citation impact. Therefore the simultaneous use of all three citation normalized counts is recommended in order to identify research strengths or specialisations.

On the other hand, the complementary use of the percentage of documents in Q1 journals is only suitable for the so-called "hard sciences" (STEM and to some extent the Social Sciences). There it can help to overcome the low significance of citation analysis for the most recent publications years, which is explicable by the extremely low time frame (Glänzel et al. 2016; Donner, 2018). This indicator cannot be applied to the Arts & Humanities due to the lack of appropriate data and the limited role of citations in scholarly communication reflected by journal and proceedings literature. Clarivate Analytics has so far remained true to the philosophy of Eugene Garfield and is still reluctant to calculate journal impact factors (JIF) in the Arts & Humanities based on data from the Web of Science Core Collection (Garfield, 1972; 2005). However, other providers of alternative journal impact measures, like Elsevier, Scimago and CWTS, have skipped this restriction. They publish their editions of CiteScore, Source Normalized Impact per Paper (SNIP) or Scimago Journal Rank (SJR) on a yearly basis, which also include the Arts & Humanities. Therefore the availability of alternative journal impact

measures seems to be of great advantage to these disciplines, as they might turn out to be crucial for evaluative practices in so far unapproachable scientific terrain. The inappropriate use of journal impact measures in evaluative contexts was already seriously criticized in the hard sciences (Glänzel & Moed, 2002). This holds all the more true for the Arts & Humanities.

The inherent limitations of journals-assignment-based classification systems, and especially the ones used in the Web of Science Core Collection, hamper the identification of research specializations. Calculated on journal level, the corresponding WoS Subject Categories are not uniquely assigned and they are very often not aligned with a University's organizational structure (but this is also true for esteem factors like subject rankings). Thus, the resulting overlap makes the interpretation of the results difficult. Furthermore, the classifications are either too general, like the ESI Categories, or too fine, like the WoS Subject Categories, to provide clear and sound information. Therefore, we consider the use of other classification systems in the future, like the modified Leuven-Budapest classification system (Glänzel, Thijs & Chi, 2016).

Both approaches proved to be useful in identifying research strengths and specializations, lead to similar results and should be used in a complementary way. We found that the greatest hindrance lies in the inherent limitations of journals-assignment-based classification systems and the considerable time and efforts for more accurate researcher-based publication analyses.

This study has also some noticeable limitations:

1) It presents a case study for only one university.
2) We compared both approaches for only two of the nine research specialisations and therefore require further validation.
3) Data coverage of the WoS Core Collection is incomplete particularly in the Social Sciences and Humanities due to the low coverage of indexed monographic literature.

However, the results already show the benefits of bibliometric analyses for the accurate definition of key-research areas and research specializations. The next iterations of their definition should include results of the current study, e.g. a more prominent inclusion of Top-WoS categories like "Physical Geography", "Optics" and "Meteorology and Atmospheric sciences". In the future, we also plan to analyse all nine research specialisations accordingly. Based on the results obtained from this case study we will define a more accurate set of research specializations. Furthermore, we plan to refine the methodology by also including alternative metrics for broader impact assessment and thereby identifying research strengths in the Humanities as well (Moed, 2017).

References

Bookstein FL, Seidler H, Fieder M, Winckler G. 2010. Too much noise in the Times Higher Education rankings. *Scientometrics,* 85, 295–299.

Burkert, GR. (2016) Forschungsgovernance durch Forschungssenate? Zu einer Neuorganisation der inneruniversitären Verantwortung durch „Innovationsrat" und Forschungsstärken. *Zeitschrift für Hochschulrecht, Hochschulmanagement und Hochschulpolitik*, Heft 4, 110-122.

Donner, P. (2018). Effect of publication month on citation impact. *Journal of Informetrics*, 12(1), 330-343.

Fumasoli, T. and Huisman, J. (2013) "Strategic Agency and System Diversity: Conceptualizing Institutional Positioning in Higher Education." Minerva 51 (2). 155-69

García, J. A., Rodriguez-Sánchez, R., Fdez-Valdivia, J., Robinson-García, N., & Torres-Salinas, D. (2013). Benchmarking research performance at the university level with information theoretic measures. *Scientometrics*, 95(1), 435-452.

Garfield, E. (1972). Citation analysis as a tool in journal evaluation: Journals can be ranked by frequency and impact of citations for science policy studies'. *Science*, 178(4060), 471–479.
Garfield, E. (2005). The Agony and the Ecstasy. The History and Meaning of the Journal Impact Factor. http://garfield.library.upenn.edu/papers/jifchicago2005.pdf
Glänzel, W., & Moed, H. F. (2002). Journal impact measures in bibliometric research. *Scientometrics*, 53(2), 171–193.
Glänzel, W., Chi, P.S., Gumpenberger, C., & Gorraiz, J. (2016). Information sources - information targets: evaluative aspects of the scientists' publication strategies. *21st International Conference on Science and Technology Indicators - STI 2016. Book of Proceedings, 708-717*. J. Woolley, Richard (Ed.). Spain: Editorial Universitat Politecnica de Valencia. http://ocs.editorial.upv.es/index.php/STI2016/STI2016/paper/viewFile/4543/2327
Glänzel, W., Thijs, B., & Chi, P. S. (2016b). The challenges to expand bibliometric studies from periodical literature to monographic literature with a new data source: the book citation index. *Scientometrics*, 109(3), 2165-2179.
Gorraiz, J., Gumpenberger, C. (2012): Bibliometric practices and activities at the University of Vienna. *Library Management*, 33 (3), 174-183.
Gorraiz, J. and Gumpenberger, C. (2015). A flexible bibliometric approach for the assessment of professorial appointments. *Scientometrics*, 105, 3, 1699-1719. DOI 10.1007/s11192-015-1703-6
Gorraiz, J. and Gumpenberger, C. (2016). Individual Bibliometric Assessment at the University of Vienna: From Numbers to Multidimensional Profiles. *El Profesional de la Información*, 25 (6), 901-914; http://arxiv.org/abs/1601.08049,
Gorraiz, J.; Wieland, M. and Gumpenberger, C. (2017). To Be Visible, or Not To Be, That Is the Question. *International Journal of Social Science and Humanity*, 7 (7), 467-471; doi: 10.18178/ijssh.2017.V7.868; http://www.ijssh.org/index.php?m=content&c=index&a=show&catid=90&id=1238
Hazelkorn E. 2007. Impact and influence of league tables and ranking systems on higher education decision-making. *Higher Education Management and Policy* 19(2), 87–110.
Huisman, J. and van Vught, F. (2009) "Diversity in European Higher Education: Historical Trends and Current Policies." In *Mapping the Higher Education Landscape. Towards a European Classication of Higher Education*, edited by Frans A. van Vught. 17-37. Dordrecht Kluwer.
Laudel, G. and Weyer, E., 2014. Where have all the scientists gone? Building research profiles at Dutch universities and its consequences for research. In *Organizational transformation and scientific change: The impact of institutional restructuring on universities and intellectual innovation* (pp. 111-140). Emerald Group Publishing Limited.
Moed, H. F. (2017). *Applied evaluative informetrics*. Springer International Publishing. ISBN: 978 3 319 60521 0
Rafols, I., Porter, A.L., & Leydesdorff, L. (2010). Science overlay maps: A new tool for research policy and library management. *J Assoc Inform Sci Technol*, 61(9), 1871-1887.
Sorz, J., Wallner, B., Seidler, H. and Fieder, M. (2015). Inconsistent year-to-year fluctuations limit the conclusiveness of global higher education rankings for university management. *PeerJ,* 3:e1217
van Vught, F. and Huisman, J., (2014). Institutional profiles: Some strategic tools. *Tuning Journal for Higher Education*, 1(1), 21-36

Detection of inappropriate types of authorship using bibliometric approaches

Nikolay A. Mazov[1] and Vadim N. Gureev[2]

[1] *MazovNA@ipgg.sbras.ru.at*
Trofimuk Institute of Petroleum Geology and Geophysics, Siberian Branch of the Russian Academy of Sciences, Koptyug ave. 3, Novosibirsk (Russia);
State Public Scientific Technological Library, Siberian Branch of the Russian Academy of Sciences, Voskhod 15, Novosibirsk (Russia)

[2] *GureyevVN@ipgg.sbras.ru.at*
Trofimuk Institute of Petroleum Geology and Geophysics, Siberian Branch of the Russian Academy of Sciences, Koptyug ave. 3, Novosibirsk (Russia);
State Public Scientific Technological Library, Siberian Branch of the Russian Academy of Sciences, Voskhod 15, Novosibirsk (Russia)

Abstract

Among a variety of types of scientists' misconduct, the problem of authorship is of special significance. Detection of such inappropriate types of authorship as guest, gift or ghost authorships implying fictive participation of one researcher or the absence in the byline of another researcher who had taken part in a study is one of the topical issues for the academic community to save and maintain integrity when carrying out research work and publishing its results. This study describes how bibliometric tools can be potentially used when detecting inappropriate types of authorship in research manuscripts. We believe that a certain distribution of publications is inherent at each stage of scientists' career progress. Significant deviations in a number of papers, co-authors, subject areas or set of journals in a certain period of work can be regarded as indicators of possible misconduct including, e.g., guest or honorary authorship. In this study, we used a set of prominent scientists of the Siberian Branch of the Russian Academy of Sciences to reveal a correlation between career progress points and unexpected increase in scholarly output possibly achieved by means of unethical co-authorship.

Introduction

A number of publications are one of the significant formal indices for assessment of the efficiency of scientists' work. Scholarly output influences the probability of grant applications to be accepted, career progress, election as academicians, and authority in scientific society. Publications are counted in case of performance review, scientific reports; a number of papers are a nearly single index when evaluating the work of young scientists (Research Metrics Guidebook, 2018). Therefore, the majority of researchers tends to publish as many as possible papers, especially in journals indexed in international scientometric databases.

Another impetus for publishing can include, for example, thesis defence which requires a fixed number of papers in some countries; work as an editor implying writing specific types of papers (e.g. prefaces, editorials, responses to readers, etc.). Dependence between career progress and indicators of scientists' efficiency lends itself to bibliometric calculation, although sociology approaches are also used for solving this issue. We believe these analyses are important when regarding publication ethics conditions which impose increasingly stricter requirements to authorship and researchers' responsibility for published results (Scott-Lichter, 2012).

This study aims to detect the possible correlation between scholarly output and career progress points including assignment to an executive position, thesis defence, and selection to the academic community (in this case the Russian Academy of Sciences). Then we analyze if publishing activity in each case corresponds to recommendations of committees on publication ethics (Defining the Role of Authors and Contributors, 2018), especially regarding inappropriate types of authorship including guest, gift, and ghost models (Rennie, Yank, & Emanuel, 1997; Yank & Rennie, 1999). To some extent, this paper continues our previous

studies on the use of bibliometrics when detecting misconducts in scholarly publishing including plagiarism detection (Mazov & Gureev, 2017; Mazov, Gureev, & Kosyakov, 2015).

Brief overview

In the course of the process of knowledge production, the number of researchers is also increasing. Change from the model "One study – one researcher – one author" to the model with multiple participants demonstrated that the system of scientific communication is not ready to give an answer concerning actual authorship in research papers and who is worth to be an author.

In the last decades, authorship phenomenon is actively discussed in medicine due to high requirements to integrity and safety of the proposed approaches to treatment. There were medical journals for which the International Committee of Medical Journal Editors (ICMJE) prepared the first Guidelines on authorship (Guidelines on authorship, 1985). These guidelines for the first time formulated criteria of authorship as follows:

- Each author should have participated sufficiently in the work to take public responsibility for the content.
- The contribution includes: (a) conception or design, or analysis and interpretation of data, or both, (b) drafting the article or revising it for critically important intellectual content, and (c) final approval of the version to be published.
- All elements of an article (a, b, and c above) critical to its main conclusions must be attributable to at least one author.
- A paper with corporate (collective) authorship must specify the key persons responsible for the article.

Besides, editors may require authors to justify the assignment of authorship. These conceptual definitions of author contribution constitute the basis for all further recommendations not only in medicine but also in other subject areas (On being a scientist, 2009; Academy of management, 2011).

With time authorship criteria have become more complicated; new practices have emerged which are not standardized according to current guidelines; the gap between the interpretation of valid and unfair authorship has been widened (Marušic et al., 2014). ICMJE criteria have been significantly specified in each subsequent version. One of the new paradigms conceptually divides a scientific article into four basic elements: ideas, work, writing, and stewardship to quantitatively evaluate the contribution of each author into those elements and to elaborate an authorship matrix (Clement, 2014). PLoS journals are a good example of partial implementation of these proposals since all authors are obliged to indicate their roles in the study which comprises 14 different positions (https://journals.plos.org/plosone/s/authorship#loc-author-contributions).

Extent detalization of authorship gave an incentive to the development of the new concept of contributors responsible to the integrity of the whole study (Rennie, Yank, & Emanuel, 1997; Smith, 1997; Rohlfing & Poline, 2012). At the same time, the model of group authorship also attracts criticism since makes equal participants of large research projects and small groups of authors. Researchers noted that addition of group authors into byline devalues author efforts, cripples stable system of evaluation of science, and opens the door for unfair authorship violating the main publishing ethical principles (Gasparyan, Ayvazyan, & Kitas, 2013; Rohlfing & Poline, 2012).

Reasons for unethical behavior are frequently connected with a deficient system of government of science, e.g., the use of formal approaches when evaluating the efficiency of researchers' work, addressing scholarly output and citations when funding, employing and promoting, etc. The topicality of protection from unfair authorship is caused by the necessity to sustain core values of scientific ethics, as well as to strengthen the image of science in society. Generally,

inappropriate authorship can be considered as a threat to the existence of science as a social institution.

To date, participants of scientific communications understand the necessity to use a complex approach to prevent inappropriate types of authorship. Particularly, the following approaches have been elaborated (Gasparyan, Ayvazyan, & Kitas, 2013):

- **Scientific organizations and universities** must implement relevant educational courses, set a policy to discourage inappropriate authorship, and develop and update policy statements and authorship criteria.
- **Publishers** must ensure proper guidance and interpretation of authorship in instructions for authors; adopt field-specific recommendations.
- **Editors** of journals must stick to authorship criteria and journal instructions, obtain author contributions statements, resolve disputes by cooperating with authors or research institutions.
- **Reviewers** must familiarize with available guidelines, report suspected authorship to editors.
- As for **authors**, they must familiarize with available authorship guidelines and journal instructions for authors; agree on the responsibilities, order and place of listing co-authors early at the start of the research, and avoid misconduct and unfair authorship by self-regulation.

Development of various methods for the detection of cases which had not been prevented is another task for the scientific and publishing community. Of note, successful detection of cases of unfair authorship is of random nature, as editors and reviewers often rely on the integrity of scientists and their adherence to author guidelines. Unfortunately, it should be acknowledged that now there are no efficient tools for the detection of violations of authorship criteria. Thus, it is recommended that strong deterrents should be established to end undeserved authorship and related fakeries (Rivera, 2019).

The bibliometric approach seems to be very promising in addressing the problem of detection of inappropriate authorship. One of the bibliometric directions includes the development of "justified" indices to evaluate the efficiency of work of a scientist. Developing of separate indices for each participant of a study reflecting his/her contribution can lead to the eradication of unfair types of authorship (Kovacs, 2013). The other study proposed to analyze the degree of incidence of the same co-authors to detect a threshold, with raising ethical concerns when it is exceeded (Bugaev, 2012). Using bibliometrics, it is possible to define typical publication behavior and typical distribution of papers for certain author connected with his/her stage of career progress and to detect significant deviations from that distribution that can raise suspicions concerning meeting the requirements of publication ethics.

Data and methods

In our study, we used a sample of prominent scientists working in research institutes of Academgorodok (Kupershtokh & Apolonskiy, 2014) belonging to the Siberian Branch of the Russian Academy of Sciences. List of organizations currently includes institutes of former Russian Academy of Medical Sciences as well as Russian Academy of Agricultural Sciences and is presented on SB RAS official website (https://www.sbras.ru/ru/organization/2134). Out of 53 institutes, we selected 39 ones since we omitted subdivisions of Novosibirsk institutes, as well as Novosibirsk subdivisions of institutes located in other Russian cities.

Our sample can be regarded as representative due to the high authority of scientists holding key positions at SB RAS institutes covering almost all scientific subject areas (Table 1). It includes 18 academicians and 6 associates of the Russian Academy of Sciences.

Table 1. Distribution of Novosibirsk research organizations according to subject areas.

Subject area	*Number of institutes*
Medicine	9
Physics	7
Chemistry	6
Life Sciences	6
Humanities and Social Sciences	6
Mathematics and Informatics	4
Earth and Planetary Sciences	3
Agricultural Sciences	2

When collecting bibliometric data, we used the Russian Science Citation Index (www.elibrary.ru) comprising more than 12 million bibliographic records of papers published by Russian authors. We analyzed all types of publications from 1988 to 2017. To reveal the publication coefficient at the moment of working as a head of the organization and before assignment to this position, we considered equal timespans. For instance, if a researcher has managed the organization for 10 years, we also analyzed his/her publications for 10 years before the date of appointment. In the case of thesis defence, we analyzed the 3-year window before and after the defence. Data on thesis defence were extracted from open electronic catalogs of the Russian State Library (http://diss.rsl.ru) and Central Scientific Medical Library (http://www.scsml.rssi.ru). When studying changes in subject areas of scientists' research, we used the State Rubricator of Scientific and Technical Information (http://grnti.ru).
We did not indicate data on specific organizations of scientists since the goal of this papers is to present the general rationale for the dependence of career progress events on publication activity and to reveal a violation of publication ethics requirements in a certain period of work. All organizations are marked with digits.

Results

Figure 1 presents the publication coefficients of the scientists from our sample in case of their assignment to leading positions. We analyzed two periods, i.e. before and after an assignment, respectively. Publication coefficient K_p^1 was calculated as follows:

$$K_p^1 = \frac{P_1}{P_1+P_2} \times 100, \quad (1)$$

where P_1 denotes the number of publications before an appointment to the leading position, P_2 is a number of papers at the time of holding an appointment. Similarly, we detected publication coefficient at the time of holding an appointment:

$$K_p^2 = \frac{P_2}{P_1+P_2} \times 100. \quad (2)$$

Multiplier 100 was included for clearer difference between the coefficients.

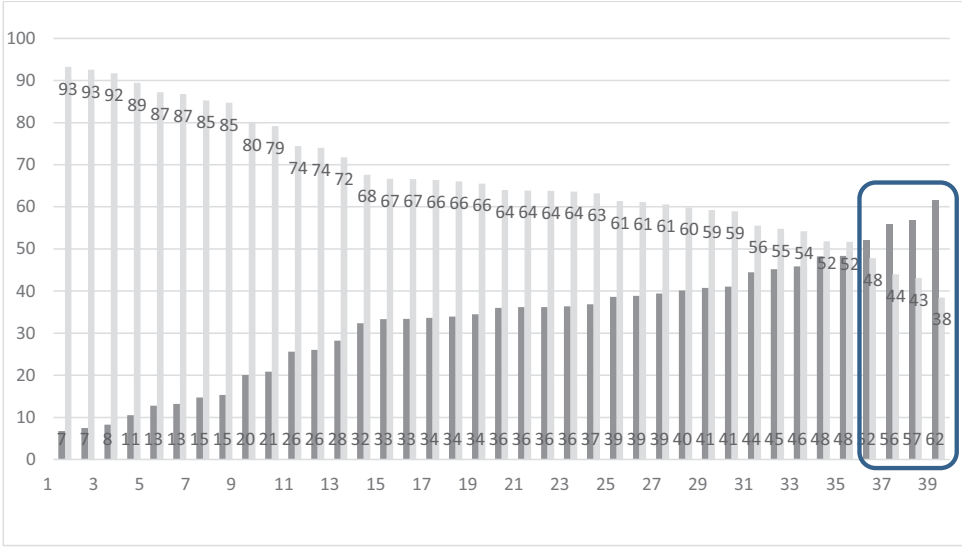


Fig. 1. Publication coefficients of researchers before and after appointment to leading positions. Dark grey denotes publication coefficient before an appointment; light grey – at the time of institute administration. Unusual cases of higher scholarly output before appointment are boxed.

Figure 1 shows that 35 scientists (90 percent) out of 39 significantly increased their scholarly output after their appointment to leading positions.

In figure 2 we revealed an indicative enhancement of a number of subject areas in publications after appointment to the leading position. Subject areas coefficients were calculated as that of publications coefficients (see formulae (1) and (2)).

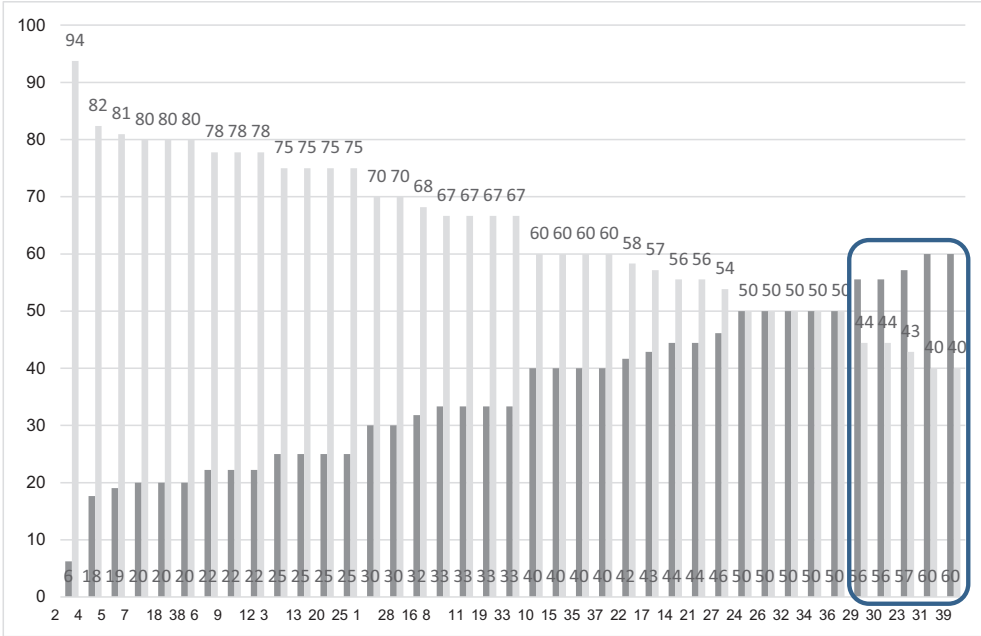


Fig. 2. Dynamics pattern of subject areas in publications by scientists before and after their appointment to leading positions in terms of the number of rubrics of the State Rubricator of Scientific and Technical Information. Dark grey denotes subject areas coefficient before an appointment to leading positions, while light grey coefficient after an appointment. Unusual cases of wider subject distributions before an assignment are boxed.

As Figure 2 indicates, only in 5 cases out of 39 subject varieties in publications decreased after assignment to leading positions, and in another 5 cases remained the same. In publications of remain 29 scientists (74 percent) topic variety in publications significantly increased.
Figure 3 demonstrates how an appointment affects the model of authorship. Co-authorship coefficient was calculated in a similar way as for publication coefficient (see formulae (1) and (2)).

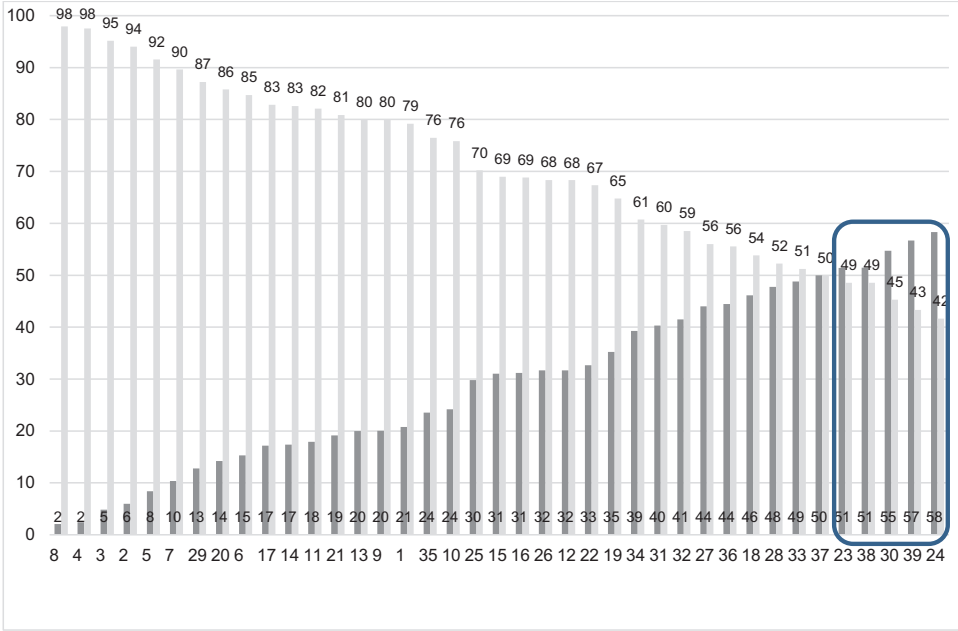


Fig. 3. The share of co-authors in papers by scientists from the sample before their appointment to leading positions and at the time of institute management. Dark grey denotes co-authors coefficient before an appointment; light grey – at the time of institute administration. Unusual cases of a higher number of co-authors before appointment are boxed.

Figure 3 shows that publications of only five researchers from our sample have a lower number of co-authors at the time of appointment, and in one case remained the same, while in most cases number of co-authors significantly increased.

Figure 4 demonstrates the distribution of papers in 3-year windows before and after an election of researchers from our sample as the members of the Russian Academy of Sciences. Coefficients were calculated according to formulae (1) and (2).

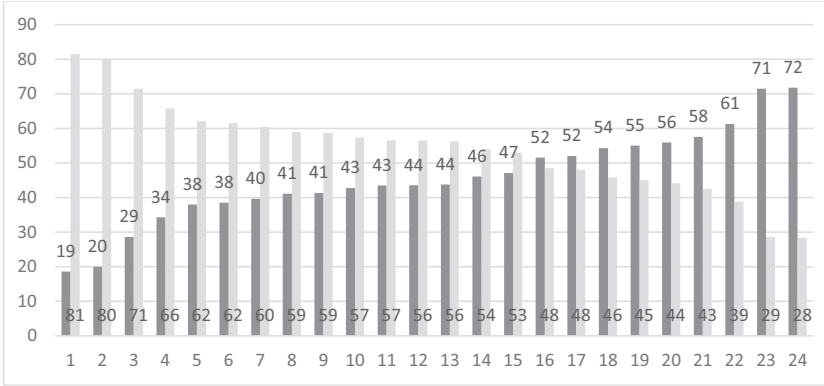


Fig 4. Publication coefficients of scientists in 3-year windows before and after the selection as members of the Russian Academy of Sciences. Dark grey stands for publication coefficient before the selection, while light grey – after the selection.

Out of 24 analyzed scientists being members of the Russian Academy of Sciences 15 researchers (63 percent) demonstrated increased scholarly output after the selection as members, while 9 scientists had decreased number of papers. Generally, except for extreme edge positions of the diagram we observe the equal distribution of papers.
Figure 5 depicts publication coefficients before and after thesis defence. Coefficients were calculated in the same manner as mentioned above.

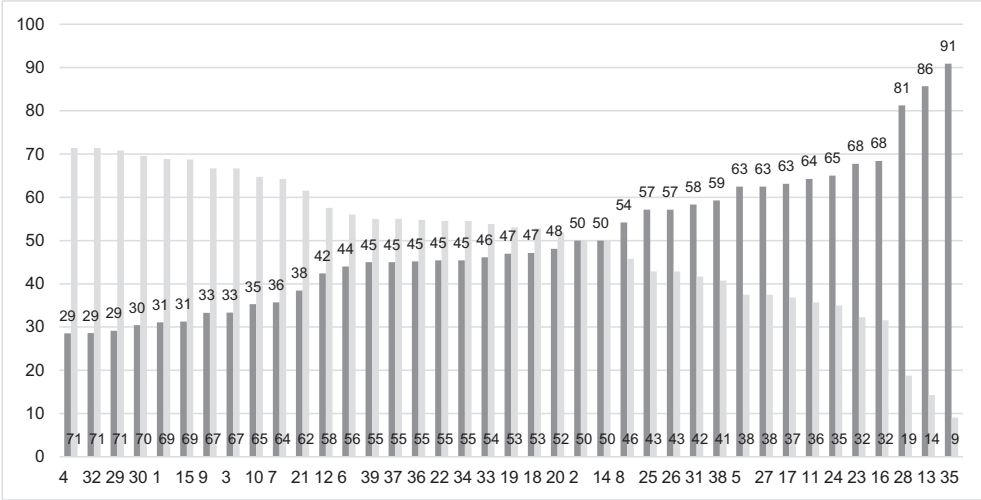


Fig. 5. Publication coefficients in 3-year windows before and after thesis defence. Dark grey stands for publication coefficient before thesis defence; light grey – after thesis defence.

Fifteen scientists decreased their publication activity after thesis defence as compared to the period of preparation of thesis; a scholarly output of 22 researchers increased, while in 2 cases remained the same.

Discussion

Using an example of Russian prominent scientists, we tried to detect dependence between changes of formal bibliometric indices and career progress events. Out of several possible cases we addressed three ones including (1) assignment to a leading position in research institute; (2) selection as a member of the Russian Academy of Sciences; and (3) thesis defence.
In all three cases, we obtained different results. The most indicative are changes in scholarly output after assignment of scientists to leading positions. Only in 4 out of 39 (Fig. 1) cases, we detected a negative trend. Furthermore, the decrease in a number of publications (right side of the diagram) was not as much expressed as rapid growth after the appointment (left side of the diagram). We believe that this growth trend is caused by the assignment to the leading position. Considering the large administrative load of scientists holding leading positions resulting in shortening the free time for research, an increase in a number of papers is achieved exclusively by means of co-authorship. As Fig. 3 depicts, it is associated mainly with the inclusion of executive in the byline as co-author. The reasons can be different including supervision in the grants, the teaching of young scientists, the inclusion of prominent name to speed peer review stages, etc.

At the same time, in the last two decades requirements to authorship have become stricter. The main criteria of authentic authorship that should be regarded together are the following (Kassirer, 1995):

- substantial contributions to conception and design, acquisition of data, or analysis and interpretation of data;
- drafting the article or revising it critically for important intellectual content;
- final approval of the version to be published;
- each listed author must be able to take public responsibility for its content.

It is important that authors must fulfil all of these criteria. Authors are also expected to designate their functional role within the group. In addition to being accountable for the parts of the work he or she has done, an author should be able to identify which of their co-authors are responsible for specific other parts of the work. In addition, an author should have confidence in the integrity of the contributions of their co-authors (Scott-Lichter, 2012). In all other cases, researchers should be mentioned in an Acknowledgment section.

A number of publishing societies have tried to prevent violation of authorship criteria, especially in the case of papers black markets selling authorship in ready-to-be-published articles (Hvistendahl, 2013). The most spread violation cases include guest, gift, and ghost types of authorship. Guest authorship has been defined as authorship based solely on an expectation that the inclusion of a particular name will improve the chances that the study will be published or increase the perceived status of the publication. The "guest" author makes no discernible contributions to the study, so this person meets none of the criteria for authorship. Honorary or gift authorship has been defined as authorship based solely on a tenuous affiliation with a study. A salient example would be "authorship" based on one's position as the head of a department in which the study took place. Ghost authors participate in the research, data analysis, and/or writing of a manuscript but are not named or disclosed in the author byline or Acknowledgments (Scott-Lichter, 2012).

Our findings enabled us to assume that assignment of scientists to leading position frequently leads to violation of publication ethics regarding international authorship criteria since the use of guest of gift authorship seems to be rampant among Russia scientists especially in the frequent absence of local standards of authorship at Russian research organizations and universities. It is confirmed by a very intensive increase in a number of papers, accompanied by an increase in a list and number of co-authors and significant enhancement of subject areas (Figs. 2 and 3). Besides, in a number of organizations especially medical ones we detected an unexpectedly high number of papers per year close to 100 items. Considering a high administrative burden, it is highly unlikely that a scientist would have enough time for publishing such a high number of papers and concurrently meet abovementioned modern requirements of authorship.

Analysis of data on the scholarly output before and after the election as the Russian Academy of Sciences members (Fig. 4) failed to detect a significant correlation between the election event and publication activity.

Distribution of papers before and after theses defense revealed that in most cases the scholarly output of analyzed scientists did not decrease after defense as one might expect, but increase. Only in 15 cases (38 percent) of 39 scholarly output decreased.

Conclusion

Authorship is considered to be one of the main sources of academic capital of a researcher. In the context of severe competition for academic positions and funding authorship is regarded to be a key index of scientific capacities and potential of a scientist (Olesen, Amin, & Mahadi, 2018). Stimulation of scientific work can lead both to loss of quality of studies and papers and to increasing a number of co-authors including cases of inappropriate authorship models.

Addressing the issue of author attribution problem should define an actual contribution of scientist to the study and divide areas of responsibilities (Gasparyan, Ayvazyan, & Kitas, 2013). Except for different author guidelines designed to provide comprehension of responsibility of researchers regarding compliance with publication ethical standards, the goal to develop technical models to detect inappropriate authorship is still topical.

In the current study, we have detected that sharp fluctuations in scholarly output can sometimes point to possible misconduct in publishing and fictive participation in the research. Especially we mean a sharp increase in a number of papers, significant fluctuation in a number and compound of co-authors, change in research areas, change in position in the byline, increase in a pool of journals with scientist's papers. Complex bibliometric analysis of a number of above mentioned and some additional changes and their comparison with certain changes in the career path of a scientist can be used for the development of powerful bibliometric tool to reveal common factors of the real and fictive contribution of scientists in publications and to develop a probable bibliometric model for detection of inappropriate types of authorship.

In further research, we plan to collect more data and to carry out more qualitative analyses including detection of thresholds at which the increase in authorships, subject areas, or co-authors would clearly indicate guest or gift authorships making it possible to develop new bibliometric model for detection of violations of publication ethics.

Acknowledgments

The reported study was funded by RFBR according to the research project № 19-011-00534. Preliminary findings were partly discussed at the 22nd International Conference and Exhibition "Information Technologies, Computer Systems and Publications for Libraries" LIBCOM-2018 (26–30 November 2018, Suzdal, Russia).

References

Academy of management code of ethics. (2011). *Academy of Management Journal, 54*(6), 1299-1306.

Bugaev, K.V. Some problems of co-authorship ethics. (2012). *Herald of Siberian institute of business and information technology, 2*(2), 72-73. [In Russian]

Clement, T.P. (2014). Authorship Matrix: A Rational Approach to Quantify Individual Contributions and Responsibilities in Multi-Author Scientific Articles. *Science and Engineering Ethics, 20*(2), 345-361. doi:10.1007/s11948-013-9454-3

Defining the Role of Authors and Contributors. (2018). Retrieved from http://www.icmje.org/recommendations/browse/roles-and-responsibilities/defining-the-role-of-authors-and-contributors.html

Gasparyan, A.Y., Ayvazyan, L., & Kitas, G.D. (2013). Authorship problems in scholarly journals: Considerations for authors, peer reviewers and editors. *Rheumatology International, 33*(2), 277-284. doi:10.1007/s00296-012-2582-2

Guidelines on authorship. (1985). International Committee of Medical Journal Editors. *British Medical Journal, 291*(6497), 722. doi:10.1136/bmj.291.6497.722

Hvistendahl, M. (2013). China's Publication Bazaar. *Science, 342*(6162), 1035–1039. doi:10.1126/science.342.6162.1035

Kassirer, J. P. (1995). Authorship criteria [3]. *Science, 268*(5212), 785-786.

Kovacs, J. (2013). Honorary authorship epidemic in scholarly publications? How the current use of citation-based evaluative metrics make (pseudo)honorary authors from honest contributors of every multi-author article. *Journal of Medical Ethics, 39*(8), 509.doi:10.1136/medethics-2012-100568

Kupershtokh, N., & Apolonskiy, A. (2014). Physics in Novosibirsk and Akademgorodok. *Physics in Perspective, 16*(2), 250-276. doi:10.1007/s00016-014-0138-4

Mazov, N., & Gureev, V. (2017). [Study results for the detection of translated plagiarism using bibliometric databases]. *Nauchnye I Tekhnicheskie Biblioteki-Scientific and Technical Libraries*(12), 87–96.

Mazov, N.A., Gureev, V.N., & Kosyakov, D.V. (2016). On the development of a plagiarism detection model based on citation analysis using a bibliographic database. *Scientific and Technical Information Processing, 43*(4), 237-241. doi:10.3103/S0147688216040092

Marušic, A., Hren, D., Mansi, B., Lineberry, N., Bhattacharya, A., Garrity, M., Clark, J., Gesell, T., Glasser, S., Gonzalez, J., Hustad, C., Lannon, M.M., Mooney, L.A., & Peña, T. (2014). Five-step authorship framework to improve transparency in disclosing contributors to industry-sponsored clinical trial publications. *BMC Medicine, 12*(1), 10 p. doi:10.1186/s12916-014-0197-z

Olesen, A., Amin, L., & Mahadi, Z. (2018). Unethical authorship practices: A qualitative study in Malaysian higher education institutions. *Developing World Bioethics, 18*(3), 271-278. doi:10.1111/dewb.12200

On being a scientist: A guide to responsible conduct in research: Third edition. (2009). Washington, DC: The National Academies Press, 82 p. doi:10.17226/12192

Rennie, D., Yank, V., & Emanuel, L. (1997). When authorship fails. A proposal to make contributors accountable. *Jama, 278*(7), 579-585.

Research Metrics Guidebook. (2018). Retrieved from https://www.elsevier.com/__data/assets/pdf_file/0020/53327/ELSV-13013-Elsevier-Research-Metrics-Book-r5-Web.pdf

Rivera, H. Fake peer review and inappropriate authorship are real evils. (2019). *Journal of Korean Medical Science, 34*(2). doi:10.3346/jkms.2019.34.e6

Rohlfing, T., & Poline, J.B. (2012). Why shared data should not be acknowledged on the author byline. *NeuroImage, 59*(4), 4189-4195. doi:10.1016/j.neuroimage.2011.09.080

Scott-Lichter, D. (2012). *CSE's White Paper on Promoting Integrity in Scientific Journal Publications, 2012 Update* (3rd Revised Edition ed.). Wheat Ridge, CO: Council of Science Editors.

Smith, R. (1997). Authorship is dying: Long live contributorship. The BMJ will publish lists of contributors and guarantors to original articles. *British Medical Journal, 315*(7110), 696.

Yank, V., & Rennie, D. (1999). Disclosure of researcher contributions: A study of original research articles in the lancet. *Annals of Internal Medicine, 130*(8), 661-670. doi:10.7326/0003-4819-130-8-199904200-00013

Impact of government intervention on publication activity: case of Russian universities

Nataliya Matveeva,[1] Ivan Sterligov[2] and Maria Yudkevich[3]

[1] *nmatveeva@hse.ru*
National Research University Higher School of Economics, 4 Slavyanskaya Square, bld. 2, 109074 Moscow (Russia)

[2] *isterligov@hse.ru*
National Research University Higher School of Economics, Myasnitskaya Str 13, 101000 Moscow (Russia)

[3] *yudkevich@hse.ru*
National Research University Higher School of Economics, Myasnitskaya Str 20, 101000 Moscow (Russia)

Abstract

Government support of the group of leading national universities, has became a common practice in many countries. The impact of such government implementation on research performance and the structure of scientific collaboration is not studied enough. We investigate the publication activity of leading Russian universities, after the Government excellence universities initiative implementation (Project 5-100). In this study, we examine quantity of changes of publication indicators, along with quality alteration of research performance. Running a linear growth model with mixed effects, we estimate the project's effect on publications (total number, per capita, publications in high-quality journals and multi-authors publications). To see the role of collaboration patterns in research output, we analyzed the changes in average numbers of authors' affiliations for publications, with a different number of authors. We reveal, that universities that got financial support, demonstrate the substantial, steady increase in publications both measured in total numbers and per capita, in comparison with universities from the control group. Although for the publications in the top 25% journals, this effect is not observed in 2016. At the same time, analysis of affiliations, allows us to demonstrate that certain results found effects are partly explained by collaborations, with other organizations, and the impact of the project in the quality of research performance, is more heterogeneous across universities.

Introduction

In recent decades, there are numerous university excellence initiatives across the world. In different countries, governments support some selected institutions in their efforts to improve performance, and enter the league of world-class universities (Möller, Schmidt, & Hornbostel, 2016; Salmi, 2015; Shin, 2009; Yonezawa & Shimmi, 2015; Zhang, Patton, & Kenney, 2013). Russia is not an exception. In Russia, the program for targeted support of the group of leading universities started in 2013, and was aimed to improve the competitiveness of Russian universities. Participating universities were chosen by competitive selection, and each year must report the progress in the number of publication and citation counts in, Web of Science (WoS) and Scopus, to secure funding for the next year. Thus, research performance has been chosen as one of the key indicators of university progress. The key feature of this program, which differentiates it from many similar excellence university programs across the world, is the short-term character of control, over performance and funding. Obviously, such a design creates strong incentives for universities for quick results, and pushes them to seek the possibility to be on a positive trend every year. In particular, T.Turko and co-authors, identified an increase in the total number of publications and the number of publications, in high-quality journals of participating universities (Turko, Bakhturin, Bagan, Poloskov, &

Gudym, 2016). In the study (Poldin, Matveeva, Sterligov, & Yudkevich, 2017), it shows that during the first two years in the 5-100 project, the total number of publications, the number of publications per capita and the number of publications in the journals of the first quartile, increased significantly. In addition, participation in the project, forced universities to change their publication strategies in favor of increasing the number of publications (Guskov, Kosyakov, & Selivanova, 2017).

At the same time, participation in the project may also push university administrations to prioritize quantity over quality, and create incentives for the faculty to publish faster, while targeting easy-to-publish, low quality journals. As we can see from presented results in A. Guskov (Guskov, Kosyakov, & Selivanova, 2018), some participating universities use unfair strategies, for example, publishing in "predatory" journals, to increase a publications' output. Analysis of affiliation, allows to see the collaboration patterns of universities, and the share of papers written in co-authorship with other organizations. Works with multiple affiliations have, on average, more citations (Sanfilippo, Hewitt, & Mackey, 2018).

So the aim of this study, is to estimate the project's impact on publication activity, and also to analyze the structure of this impact: whether this program results in greater quantitative, and qualitative output. To do so, we look at the relative growth of research output in 5-100 universities, compared to the output of the control group universities, which were not included in the first wave of the project. Namely, we investigate the impact of the project on all publications, publications in different journal quartiles and multi-authors publications. To understand the cause of the results, we look at the collaboration patterns of universities, by analyzing the dynamics of authors' affiliations.

Why is the case of Russia especially interesting?[1] Russia represents a mature academic system, with many disciplines having produced research at the highest level, although not always visible in WoS or Scopus, due to several reasons (Moed, Markusova, & Akoev, 2018). At the same time, for many decades of the Soviet period, there was a clear divide between academia (represented by numerous research, non-teaching institutions coordinated by the Russian academy of sciences), and higher education institutions (mostly focused a teaching mission). Such reforms and their quick implementation, have a profound impact on the internal structure of universities, faculty contracts, salaries, career concerns of academics and many other aspects of a universities life. In most cases, the ambitious goals of boosting international research performance, are considered by a faculty, as externally imposed by the university administration and public authorities, and sometimes, are not supported by existing academic norms. Under these conditions, the need for recording research performance information, both data and administrative aspects, is especially relevant.

Data

Our sample consists of 14 universities, that participated in Project 5-100, since 2013 (treatment group[2]), and 13 other Russian universities (control group)[3]. The control group

---

[1] For a critical review of modern Russian state research and innovation policy see (Dezhina, 2017)

[2] Far Eastern Federal University (FEFU), Kazan Federal University (KFU), Moscow Institute of Physics and Technology (MIPT), National University of Science & Technology (MISIS), National Research Tomsk State University (TSU), National Research Tomsk Polytechnic University (TPU), National Research Nuclear University (MEPhI), Lobachevsky State University of Nizhny Novgorod (UNN), Novosibirsk State University (NSU), Samara National Research University (SSAU), St. Petersburg State Polytechnical University (SPbGPU), St. Petersburg State Electrotechnical University (LETI), St. Petersburg State University of Information Technologies (ITMO), Ural Federal University (UFU).

includes universities, that at the beginning of the project, had relatively comparable key indicators of research scale and intensity, with those of the treatment group: number of publications and number of publications in highly cited journals. Here we have to admit that these indicators were, on average, somewhat lower than those of the 5-100 universities, while we included those with a minimal gap from the treatment group. We also aimed to choose universities with similar disciplinary profiles, and thus excluded medical universities.

We used data regarding the total number of journal articles and reviews, from 2010 to 2016, attributed to universities' profiles in WoS (indexes SCI-expanded and SSCI, document types "article" and "review"), and number of publications in the journals of the highest (Q1), and the lowest (Q4) quartiles, according to their Journal Impact Factor (JIF; using quartiles instead of rough JIF values, mainly removes the problem of varying journal citation levels across different subject areas). We also collected data regarding the number of scientific staff at each university, and the amount of R&D funding from the statistics depository, of the Ministry of Education and Science of Russia.

The number of publications has increased in both groups of universities in the analyzed period, although the pace of growth differs. The gap between the treatment and control group in funding and number of publications has become wider since 2013. So before program implementation, total number of publications, number of publications per capita, amount of funding per capita in the 5-100 universities, were approximately twice as high in the control group. At the same time, during this period, the control group has a higher number of scientific staff than 5-100 universities.

Methods

To estimate effect of the 5-100 project and variation of the effect by years, we use a linear mixed-effects model (LME) with correlated random trend. This model is a generalized type of linear regression model, and it allows to take into account variation of parameters over time, and over individuals (Verbeke, 1997). Estimated effects may be both fixed and random. The fixed part is an observed relation between variables represented in regression coefficient, and random part is an unobserved deviation. To take into account individual dynamics between universities, we used a LME model with correlated random trend (Wooldridge, 2002). The model is applicable in the case of presence of individual growth trends in the sample.

The basic model is presented by Eq.(1):

$$I\ (p\quad n\quad n\ i)_i = \alpha_0 + \alpha_i b + (\beta_0 + \beta_i)\cdot l(y\quad)_t i + \lambda_t ee\quad aa\quad rt$$
$$+\delta^2\quad\cdot d(y\quad = 2\quad e)_t{}^4 + \delta^2\ 0\ \alpha d(y\quad = 12\ r\ e) + \delta^2\ 40\cdot d(y\quad = 2\ r\ e) + \xi_i \quad (1)$$

α_i, β_i – random intercept and trend coefficients (not estimated);
α_0, β_0 - fixed intercept and trend coefficients;
$(year)_t$ – time trend (… -3, -2, -1, 0, 1, 2, 3…);
λ_t – time dummies (for taking into account control group);
δ - average treatment effect, ATE;
d_t-dummy for project years;
d_i-dummy for participants;

---

[3] Baltic Federal University (BFU), North-Eastern Federal University in Yakutsk (NEFU), Peoples Friendship University of Russia (RUDN), Siberian Federal University (SibFU), Tyumen State University (TyUU), South Ural State University (SUSU), Moscow Aviation Institute (MAI), Perm National Research Polytechnic University (PGTU), Saratov State University (SSU), Southern Federal University (SFU), Bauman Moscow State Technical University (MSTU), Voronezh State University (VSU), Ufa State Aviation Technical University (UGATU).

ε_{it}- standard errors;
Variable covariates have been added: for total number of publications, covariates are R&D funding and number of scientific staff, for number of publications per capita as a response, variable covariate is R&D funding per capita.
To estimate the 5-100 project effect total number of publications, number of publications in Q1 journal, number of publications in Q4 journals, number of multi-author publications and their normalized values by number of scientific staff, were used as the response variable.
Multi-author works were chosen, due to analysis of the Pearson correlation for publication dynamics in 2009-2016. The values of correlation between all publication dynamics, and dynamic of publications with certain number of authors, decrease after ten authors. Thus, we use publications with ten and more co-authors as multi-author works. For the 5-100 universities, 10% of all publications have 11 and more authors, and for control group universities, this value is about 2 %.
To investigate the effect of the project on university collaboration patterns, we analyze the dynamics of affiliations by authors in the treatment and the control group. The WoS publication records were filtered by number of affiliations, by number of authors and by journal quartiles. The affiliations-to-author ratio, was calculated taking into account the subtraction of the minimum value of this ratio, which corresponded to the number of single affiliations for any number of authors. For example, for publications with one affiliation, the ratio of affiliations/authors can't be less than 1. For publications with one author, the subtraction is 1/1 (the ratio is 1-1/1). For publications with two authors is 1/2 (the ratio is 1-1/2), and so on for any number of authors.

Findings

Participation in the 5-100 Project, positively affects the number of publications (see Table 1 for results of regression analysis). The greatest effect, is observed in the second year of participation, and the number decreased in 2016. Estimates are calculated in exponential form, in the tables their linear modifications are presented. For instance, coefficient 1.352 for the variable *Year=2014×participant of 5-100* in first column, means that in this model specification, the 5-100 Project universities, outperform the general publications trend by 35.2%, in 2014. If the value of the coefficient is less than one, the growth of the dependent variable, is due to the reduction of the response variable.
We consider 4 specifications of the model: specifications 1-2 are for full sample, and specifications 3-4 are for the treatment group only. This separation shows, what effect on participating universities exists, in comparison with the general publication trend, and how these universities outperform their own trend. As we can see from Table 1, the values of the effect are slightly higher for the full sample.

Table 1. The effect of the Project for total number of publications

	(1)	(2)	(3)	(4)
Year=2014×participant of 5-100	1.352*** (0.078)	1.367*** (0.084)	1.273*** (0.082)	1.252*** (0.076)
Year=2015×participant of 5-100	1.578*** (0.141)	1.588*** (0.148)	1.456*** (0.147)	1.441*** (0.143)
Year=2016×participant of 5-100	1.523*** (0.145)	1.512*** (0.145)	1.353*** (0.154)	1.367*** (0.161)
Year (β_0)	1.186*** (0.028)	1.196*** (0.029)	1.196*** (0.030)	1.178*** (0.035)
Scientific staff (thousand)	1.099* (0.058)	1.109* (0.065)	1.011 (0.040)	1.006 (0.033)

Funding of R&D (bln. rubles)		0.932 (0.072)		1.106 (0.102)
Control group	yes	yes	no	no
Time effects	yes	yes	no	no
N	189	189	98	98

Standard errors in parentheses
$^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$

To understand how the project affects productivity of scientific staff, we estimate the effect on publications per capita. The effect is positive, and its value varies depending on the model specification. With taking into account the control group, the values of the effect are higher. That is, participating universities enhance the general trend by a larger value than their own. This effect grew the entire project (from 35.4% in 2014 to 69.4% in 2016), but this growth can partly be explained by a reduction of scientific staff, in 2016.

Thus we found that participating in the project, allowed universities to surpass general publication activity, more than 35% in 2014, and more than 50% in 2016. However, to be able to conclude whether the project was successful, one needs to understand the project influence on publications of different quality. So we then look at the relative growth of high-quality (Q1) publications, in comparison with Q4 output.

In both groups, the number of publications and number of publications per capita in the Q1 journals, has increased, but the growth rate is higher in the 5-100 group. In 5-100 universities, number of publications in Q1 journals are much higher than in the control group In 2014, participating universities increased the number of their publications in Q1 by 74%, and in the following years the rate of growth slowed down. In the control group, the total number of publications in Q1 journals, increased monotonically, in the analyzed period.

The positive significant effect of the project on high-quality publications, is observed in the first two years (Table 2). Then, the number declined in 2015 (43% in 2014 and 33% in 2015). The presented results, also provide for two specifications, which differ in the number of explanatory variables. The effect of 5-100 Project on the number of publications in Q1 journals per capita, was detected in 2014 and 2015. The highest number is observed in 2014 (44.6%). Funding per capita, positively correlate with the number of publications in Q1 journals per person.

Table 2. Results for total number of publications in Q1 and for number of publications in Q1 per capita

	Q1	*Q1*	*Q1 per capita*	*Q1 per capita*
Year=2014×participant of 5-100	1.433^{***} (0.144)	1.434^{***} (0.164)	1.446^{***} (0.146)	1.367^{***} (0.154)
Year=2015×participant of 5-100	1.328^{*} (0.202)	1.332^{*} (0.213)	1.322^{*} (0.220)	1.332^{*} (0.197)
Year=2016×participant of 5-100	1.119 (0.182)	1.139 (0.185)	1.283 (0.240)	1.282 (0.222)
Year (β_0)	1.346^{***} (0.043)	1.341^{***} (0.045)	1.339^{***} (0.042)	1.282^{***} (0.040)
Scientific staff (thousand)		1.043 (0.111)		
Funding of R&D (bln. rubles)		1.022 (0.180)		
Funding of R&D per capita (mln. rubles)				1.688^{**} (0.363)
N	182	182	182	182

Standard errors in parentheses
$^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$

The government controls general output, without differentiation of different quality segments, so to maximize, it universities may be interested in increasing low-quality output (Guskov, Kosyakov and Selivanova, 2018), which is often cheaper, easier, and faster. So we look at the relative dynamics of Q4-output. The project has a positive significant effect on Q4 publications in all three years (Table 3). Participating universities outperform the general trend of publications in Q4, more than 23% in 2014, and more than 42% in 2016. Taking into account the number of scientific staff, the highest number is found in 2016 (48.9%). The significant effect of the project on the number of publications in Q4 per capita, was also revealed in all years. The highest number is in 2016 (more than 50.3%). For the number of publications in Q1 journals, we observed completely opposite results - the highest number is found in 2014 and then decreased in 2016.

Table 3. Results for total number of publications in Q4 and for number of publications in Q4 per capita

	Q4	*Q4*	*Q4 per capita*	*Q4 per capita*
Year=2014×participant of 5-100	1.239***	1.257***	1.204**	1.167*
	(0.079)	(0.089)	(0.100)	(0.110)
Year=2015×participant of 5-100	1.456***	1.477***	1.375***	1.363***
	(0.111)	(0.122)	(0.138)	(0.138)
Year=2016×participant of 5-100	1.425***	1.489***	1.527***	1.503***
	(0.114)	(0.124)	(0.176)	(0.182)
Year (β_0)	1.120***	1.122***	1.121***	1.103***
	(0.024)	(0.024)	(0.024)	(0.025)
Scientific staff (thousand)		1.195***		
		(0.062)		
Funding of R&D (bln. rubles)		0.965		
		(0.065)		
Funding of R&D per capita (mln. rubles)				1.226*
				(0.135)
N	189	189	189	189

Standard errors in parentheses
$^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$

Another way to investigate the quality of publication output, is to analyze the publications with many authors. These works assume a special form of collaborations, with minimal contribution per author. Although on average, that works have 35% of citations, from citations of all publications of the treatment group, in 2012-2016. For NSU, this indicator is 63%. For MEPI 81%. For most of the mentioned universities, publications with ten and more co-authors, are highly cited, and are based on experiments in Mega-science installations in high-energy physics. During the observed period, the growth of these works is typical for the selected 5-100 universities, which have experience in relevant fields (MEPhI, NSU, MIPT, SPbGPU, TSU).

As seen in Table 4, the effect of the project on these works is huge, both for total number of these publications, and for their normalized values. The value has increased from 2014 to 2016. The coefficient 2.726, in the first column of Table 4, means that 5-100 universities outperform the general trend by 272%, and in 2016 by 465%. The number of scientific staff negatively correlates with the number of such publications. Funding per capita positively correlates with multi-authors works per capita.

Table 4. Results for multi-authors publications

	Total	*Total*	*Per capita*	*Per capita*
Year=2014×participant of 5-100	2.726***	2.771***	2.646***	2.326***
	(0.487)	(0.534)	(0.491)	(0.426)
Year=2015×participant of 5-100	3.536***	3.553***	3.515***	3.128***
	(0.929)	(0.954)	(0.936)	(0.763)
Year=2016×participant of 5-100	4.651***	4.419***	5.165***	4.364***
	(1.328)	(1.262)	(1.594)	(1.257)
Year (β_0)	1.048	1.053	1.047	1.019
	(0.049)	(0.063)	(0.051)	(0.057)
Scientific staff (thousand)		0.746*	2.646***	1.690*
		(0.114)	(0.491)	(0.472)
Funding of R&D (bln. rubles)		0.943		
		(0.230)		
Funding of R&D per capita (mln. rubles)				2.326***
				(0.426)
N	163	163	163	163

Standard errors in parentheses
$^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$

A common strategy to quickly increase publication output, is to increase cooperation with other higher education institutions and research organizations, in Russia and abroad. Does this really take place?

The first thing that one can easily see, is that scholars from 5-100 universities publish fewer and fewer single-affiliation papers, while in the control group this share is relatively stable and even grew in 2016 (see Fig.1). That is, in participating universities the share of publications written in co-authorship with other organizations, increased since 2012.

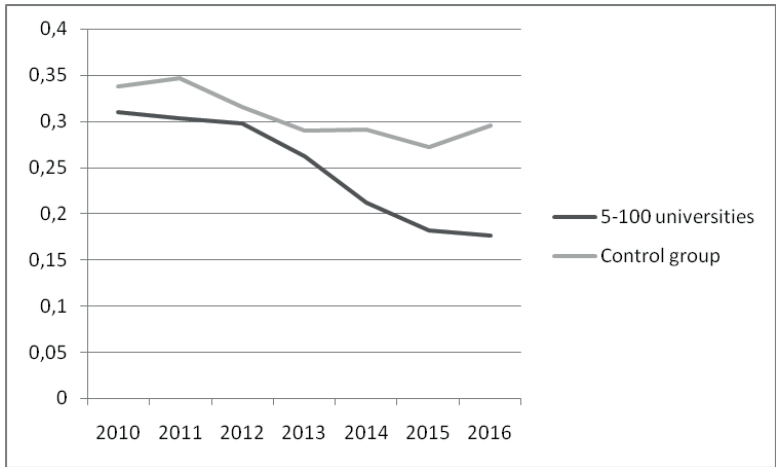


Figure 1. Share of publications with single affiliation by years

Next, we analyzed the changes in the average number of authors' affiliations for publications, with a different number of authors. The division by number of authors, allows to determine two different patterns of research collaboration in the 5-100 universities. The first, is increasing the number of affiliations of one author since 2013 (Fig. 2). This reveals that after joining the project, 5-100 universities increased the number of publications, partly due to

work of scientists, who have multiple affiliations, i.e. work in different organizations outside the 5-100. The number of affiliations for works with 2-4 authors, also increased since 2013. That demonstrates the increase of papers which were prepared in collaboration with other organizations. In universities of the control group, the stable increase of number of affiliations by author, isn't observed until 2016, when the number of affiliation by author for publications with 1 author, has increased rather drastically.

For the high-quality publications, the same trend is observed. Participating universities in Q1 journals, demonstrate a stable growth of affiliations for publications with 1-2 authors, since 2013. Moreover, the number of affiliations by one author in Q1, is noticeably higher than that in all types of work and in the Q4. The control group universities demonstrate, both increase and decrease of this indicator in the observed period. Scientists from the control group have less affiliations in Q1, than scientists from the 5-100 universities. The 5-100 universities also demonstrate a growth of authors affiliations in publications in Q4. In the period 2013-2015, the number of affiliations of one author has increased about two times. In 2016 these values decreased. On the contrary, in the control group, the number of authors affiliation has increased dramatically for publications with one author, in 2016.

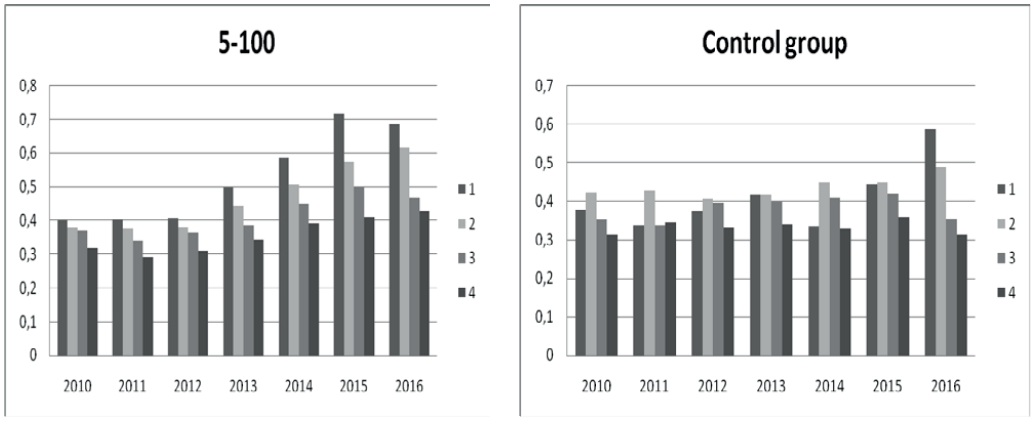


Figure 2. Number of affiliations by authors for publications with 1 - 4 authors

Since publication dynamics are different in Q1 and Q4 segments. We assumed the patterns of collaborations in these segments can also be different. Before 2014 researchers from control group universities have more affiliations than those from 5-100 universities, especially in Q1. Since 2012-2013, in 5-100 universities, the growth of authors' affiliations is observed both in Q1, and in Q4 publications. The gap between the two group of universities decreases, and in 2016 the average number of author affiliations of the 5-100 universities, is about the same as in the control group.

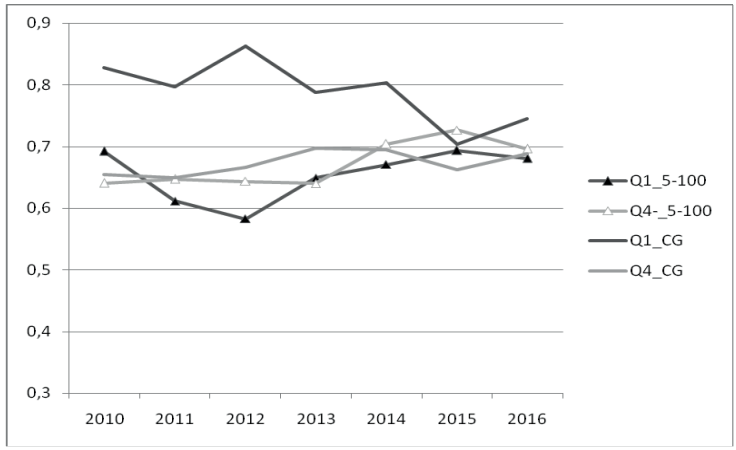


Figure 3. Dynamics of average number of author affiliations by journal quartiles

Thus since 2013, participating universities intensified not only "traditional" collaboration, when academics from different organizations joint their resources to produce quality work, but also increased "formal" collaboration between organizations, which consists in the multiple affiliations of one researcher. More often researchers from the 5-100 universities have double affiliations in Q1 publications. For example, in 2016 one researcher, on average, had almost two affiliations in Q1, and 0.6 affiliations in Q4 publications.

We also examine the share of these publications in Q1 and Q4. For each year, were calculated percent of works with one author, and the number of different affiliations from all works with one author in a year.

As shown in Fig.4, in Q1 in the 5-100 universities, since 2012 the share of publications, which one author has one affiliations, dropped greatly. The share of publications with 2 affiliations per author, has symmetrical opposite dynamics (that is, more often one author increased his number of affiliations, by up to 2 affiliations). Since 2013 the percent of publications with two affiliations by one author, is higher than percent of publications with one affiliations. The percent of works with 3-4 affiliations has also increased since 2013.

In the control group, the share of publications in Q1 with one author and one affiliation, dropped in 2013-2015. For the publications with 2 affiliations by author, there is an

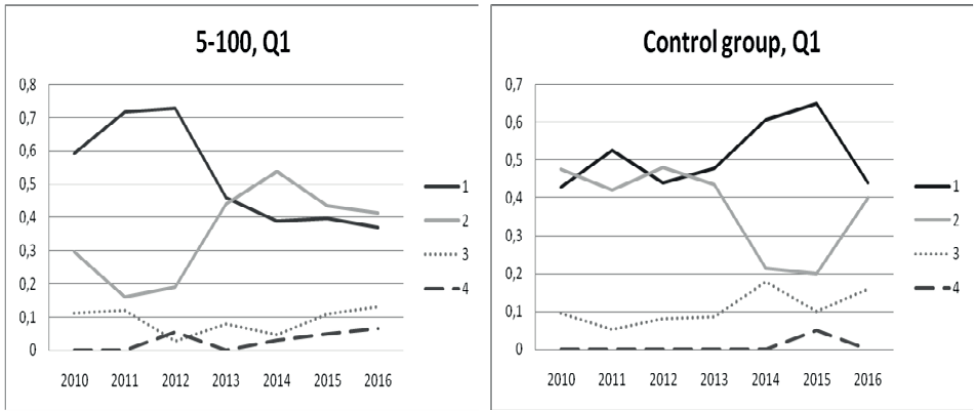


Figure 4. Share of publications with one author and 1-4 affiliations in Q1

symmetrical opposite dynamics.
As for Q4, in the 5-100 universities the share of publications with one affiliation and one author decreased in 2011-2013 and in 2014-2016. However during all observed period, single-authored publications from 5-100 universities mostly had one affiliation. The growth of affiliations per solo author, is observed in 2014 (up to 2 affiliations by one author) and in 2015 (up to 3-4 affiliations by one author). For the control group in Q4 journals, the share of publications with one affiliation by one author decreased in 2013-2015, and the share of works with 2 affiliations increased in 2013-2014 almost symmetrically.
Finally, we study the dynamics of papers written in international collaboration. In 2010, the 5-100 universities and control group have unequal share of publications, which were written in co-authorship with foreign colleagues. Then, in the 5-100 universities, this indicator increased from 33% in 2012 to 44% in 2016. The control group intensified international collaboration in 2011-2013, then the share of such publications was stable.

Discussions and conclusions

Our analysis shows that the 5-100 Project has a significant effect on participating universities. We demonstrate that this growth is expressed both in quantitative terms, and in terms of the structure of the research output. A regression analysis of the effect of the project on the total number of publications, showed that there is a positive effect of the project over the three years. Participating universities have surpassed their own indicators less than the general ones (with taking into account control group). The largest movement on the general trend, is observed in 2015, 58%, the smallest is in 2014, 36%. With taking into account the change in the number of scientific staff the greatest increased is observed in 2016, 69%.
The positive effect of the project on the total number of publications, and the number of publications per capita in Q1 journals, is observed in 2014 and 2015. The lowest values are in 2015. There is an opposite dynamics for the total number of publications, and the number of publications per capita in Q4 journals: the values are significant and increase during three years.
We also found a significant effect on multi-authors works: from 200% in 2014 up to 400% in 2016. Such publications usually include works that have a natural-science orientation and are often produced at international facilities.
In addition, we show that collaboration patterns of the universities change in the course of the project, having a qualitative impact on the general research output of the university. In 5-100 universities, the share of papers written in co-authorship with other organizations, has greatly increased since 2013. After joining the project, universities intensified their collaboration with other scientific organizations, which is reflected in a significant increase of joint publications between different organizations, as well as an in increase the number of affiliations per person. At the same time, the growth of publications of the 5-100 universities, is at least partly achieved by increasing the number of publications in Q4 journals, or by involving authors who continue to work in other organizations. That strategy is reasonable at the initial stage, and can improve scientific collaboration. However, it is less likely to be the main driver of 5-100 success, because with taking into account the project's goals, it would be more logical to produce more publications of the core faculty, in highly-cited journals.
Our findings underline the highly problematic nature of excellence initiatives, based on rankings and formal scientometric assessments: on the one hand, there is a marked and rapid increase of publication output, including the growth of papers in highly-cited and highly selective international journals. On the other hand, the scope and speed of this same increase, means that by no means was the majority of research leading to these papers prepared, using 5-100 funds. Nowadays even the publication of already prepared manuscript, can take several years, especially in top journals in some disciplines (Bjork & Solomon, 2013) in addition to at

least 1-2 years needed for every scientific project's design and completion. This means that the project's real success, at the early stage, was mostly in adding universities' affiliations to the papers prepared elsewhere, with the help of authors with multiple affiliations. Such a specific collaboration pattern, primarily aimed a quick increase of bibliometric KPIs and ranking positions (Bornmann & Bauer, 2015), is becoming more and more widespread, as we see a rapid increase in the share of authors with multiple affiliations, in control group universities in the most recent year observed. Increased collaboration, which is at least partly driven by the global advent of formalized evaluation regimes (Dahler-Larsen, 2015), means that it is becoming more and more difficult to assess individual organizations, using standard bibliometric apparatus employed by funders and governments.

Acknowledgments

The work was prepared within the framework of the HSE University Basic Research Program and funded by the Russian Academic Excellence Project '5-100'

References

Bjork, B. C., & Solomon, D. (2013). The publishing delay in scholarly peer-reviewed journals. *Journal of Informetrics*, *7*(4), 914–923. https://doi.org/10.1016/j.joi.2013.09.001

Bornmann, L., & Bauer, J. (2015). Which of the world's institutions employ the most highly cited researchers? An analysis of the data from highlycited.com. *Journal of the Association for Information Science and Technology*, *66*(10), 2146–2148. https://doi.org/10.1002/asi.23396

Dahler-Larsen, P. (2015). Spaziofilosofico 1/2015.

Dezhina, I. (2017). Science and innovation policy of the Russian government: A variety of instruments with uncertain outcomes. *Public Administration Issues*, *26*(2017), 7–26. https://doi.org/10.17323/1999-5431-2017-0-5-7-26.7

Guskov, Kosyakov and Selivanova, I. V. (2017). Strategies to improve publication activities of the universities participating in Project 5-100. *Nauchnye I Tekhnicheskie Biblioteki-Scientific and Technical Libraries*, 5–18.

Guskov, Kosyakov and Selivanova, I. V. (2018). Boosting research productivity in top Russian universities : the circumstances of breakthrough. *Scientometrics*, *8*. https://doi.org/10.1007/s11192-018-2890-8

Moed, H. F., Markusova, V., & Akoev, M. (2018). Trends in Russian research output indexed in Scopus and Web of Science. *Scientometrics*, *116*(2), 1153–1180. https://doi.org/10.1007/s11192-018-2769-8

Möller, T., Schmidt, M., & Hornbostel, S. (2016). Assessing the effects of the German Excellence Initiative with bibliometric methods. *Scientometrics*, *109*(3), 2217–2239. https://doi.org/10.1007/s11192-016-2090-3

Poldin, O., Matveeva, N., Sterligov, I., & Yudkevich, M. (2017). Publication Activities of Russian Universities: The Effects of Project 5–100 Oleg. *Educational Studies Moscow*, (2), 8-35.

Salmi, J. (2015, November). Excellence Initiatives and World-Class Universities, (Vol.6), 17-18.

Sanfilippo, P., Hewitt, A. W., & Mackey, D. A. (2018). Plurality in multi-disciplinary research : multiple institutional affiliations are associated with increased citations, 2–9. https://doi.org/10.7717/peerj.5664

Shin, J. C. (2009). Building world-class research university: The Brain Korea 21 project.

Higher Education, *58*(5), 669–688. https://doi.org/10.1007/s10734-009-9219-8
Turko, T., Bakhturin, G., Bagan, V., Poloskov, S., & Gudym, D. (2016). Influence of the program "5–top 100" on the publication activity of Russian universities. *Scientometrics*, *109*(2), 769–782. https://doi.org/10.1007/s11192-016-2060-9
Verbeke, G. (1997). Linear Mixed Models for Longitudinal Data, (October 2014), 63–153. https://doi.org/10.1007/978-1-4612-2294-1_3
Wooldridge, J. M. (2002). Econometric Analysis of Cross Section and Panel Data. *Humor - International Journal of Humor Research*, *16*(4), 752. https://doi.org/10.1515/humr.2003.021
Yonezawa, A., & Shimmi, Y. (2015). internationalization : challenges for top universities and government policies in Japan. *Higher Education*, *70*(2), 173–186. https://doi.org/10.1007/s10734-015-9863-0
Zhang, H., Patton, D., & Kenney, M. (2013). Building global-class universities: Assessing the impact of the 985 Project. *Research Policy*, *42*(3), 765–775. https://doi.org/10.1016/j.respol.2012.10.003

Participation of 'international national organisations' in Africa's research: A bibliometric analysis of two research fields in Zimbabwe

Similo Ngwenya[1] and Nelius Boshoff[2]

[1] *sngwenya@sun.ac.za*
Centre for Research on Evaluation, Science and Technology (CREST) and DST-NRF Centre of Excellence in Scientometrics and Science, Technology and Innovation Policy (SciSTIP), Stellenbosch University (South Africa)

[2] *scb@sun.ac.za*
Centre for Research on Evaluation, Science and Technology (CREST) and DST-NRF Centre of Excellence in Scientometrics and Science, Technology and Innovation Policy (SciSTIP), Stellenbosch University (South Africa)

Abstract

The study investigated the participation of so-called 'international national organisations' (INOs) in agricultural and health research in Zimbabwe, a country in southern Africa. An INO refers to an international organisation with an African country address or an initiative of an international organisation (or set of organisations) that has an African country address. A first objective was to develop a classification framework of authorship types that differentiates co-affiliation from co-authorship while also accommodating the phenomenon of INOs as a form of international participation next to international co-authorship. A second objective was to apply the framework to the research output of Zimbabwe during the period 1980–2016, to determine whether changes in authorship types coincide with changes in the country's socio-economic climate. The dataset comprised 11,056 articles and was compiled by integrating relevant articles from three sources: Scopus, Web of Science and the National Research Database of Zimbabwe. The results showed that 36% of articles in agriculture involved an INO in the period 2009–2016. The corresponding figure for health was 15%. Participation by INOs rarely occurred without any international co-authorship also being simultaneously present. The study showed that relatively small tailor-made bibliometric datasets, developed for African countries with small science systems, have the potential to produce new insights and frameworks to direct future studies on the African research landscape.

Introduction

The last decade saw a number of bibliometric studies that highlight the extent of international participation in Africa's research (e.g. Adams, Gurbey & Hook, 2014; Boshoff, 2009, 2010; Landini, Malerba & Mavilia, 2015; Mêgnigbêto, 2013; Owusu-Nimo & Boshoff, 2017; Pouris & Ho, 2014). The studies typically focus on international collaboration as one form of international participation, using analyses of international co-authorship. Recently, the inclusion of funding acknowledgments in the major bibliometric databases provided another way of measuring international participation in Africa's research (Kozma, Medina & Costas, 2018). A less obvious and often overlooked form of international participation in the African research landscape is the activity of so-called 'international national organisations' (INOs). An INO refers to an international organisation with an African country address or an initiative of an international organisation (or set of organisations) that has an African country address. An INO, therefore, uses the address of the African host country in its publications and, to some extent, tends to be adapted to the host country, following its research goals and research agendas.

In the present study, the focus is on the participation of INOs in the broad fields of agricultural and health research in Zimbabwe. Agriculture and health have always been a key focus of research among African countries (AU-NEPAD, 2010). This is the case because of the need for African countries to deal with calamities such as widespread diseases and pandemics, food security and drought. It should also be noted that current bibliometric practice commonly analyses instances of international co-authorship as a form of international participation, most probably because of ease of measurement. A new focus on INOs thus has the potential to enrich current bibliometric practice.

In light of the above, this study had two objectives. The first objective was to develop a classification framework of authorship types that accommodates the phenomenon of INOs as a form of international participation next to international co-authorship. The second objective was to apply the framework to the research output of Zimbabwe during the period 1980–2016, to determine whether changes in the authorship types (specifically authorship types that involve INOs and international co-authorship) coincide with changes in the country's socio-economic climate.

This resulted in the following research questions being formulated:

- What changes are observed when applying a new classification framework of authorship types to agricultural and health research in Zimbabwe for the period 1980–2016?
- How do the changes correspond with changes in the country's socio-economic environment between 1980 and 2016?
- What is the contribution of INOs to agricultural and health research in Zimbabwe during the period 1980–2016?
- To what extent does the contribution of INOs to agricultural and health research in Zimbabwe coincide with international co-authorship?

Before addressing the objectives and research questions of the study, a brief profile of the research and development (R&D) landscape of Zimbabwe is first presented, together with an overview of the socio-economic climate of Zimbabwe in the period 1980–2016. The overview is important because bibliometric analyses are best understood within a country's unique context (Sugimoto and Lariviere 2018).

R&D profile of Zimbabwe

As with other countries in the African region, Zimbabwe struggles to reach a recommended target of allocating 1% of its Gross Domestic Product (GDP) to R&D. The last known national survey on R&D for Zimbabwe was conducted in 2012. However, the survey was limited in that it only comprised the government and higher education sectors. Any estimates are therefore conspicuous. Based on this limited information, it was estimated that gross expenditure on R&D (GERD) represented 0.76% of GDP in 2012 (UNESCO, 2014). In that same year (2012), the country had 2,739 headcount researchers (2,511 researchers in the higher education sector and 228 in the government sector). The total number of full time equivalent researchers was 1,315. Of this total, the majority (1,205) were in the higher education sector, followed by the government sector (109).

Socio-economic overview of Zimbabwe, 1980–2016

Following independence in 1980, the country has undergone a series of socio and economic challenges that adversely affected all sectors of society, including human capital and R&D. The relevant events can be grouped into four distinct periods (1980–1990, 1991–1997, 1998–2008 and 2009–2016) and are discussed as such.

1980–1990: Vibrant economy, neo-socialist policies and brain drain

Post independent Zimbabwe, in the first years immediately after 1980, had a vibrant economy (UNESCO, 2014). The production capacity of most sectors of the economy responded positively to the advent of independence where war expenses were channelled towards economic growth in a time of peace. However, this growth only lasted into a few years of independence. The government engaged in experimental neo-socialist policies like free education, minimum wages, free low cost of housing, free health, and many more. These policies strained the economy that had been structured to provide such to only the white community. Soon government expenditure incrementally outstripped fiscal revenue and key sectors began to bulge, namely agriculture, mining and tourism in their order of importance to

the national economy. By the end of the 1980s, the economy was showing signs of stress owing to misappropriated government expenditure (Chimboza, 2012; Zvobgo, 2003).

Also of importance to note is that, in the first years of independence, up to 20,000 skilled and professional Europeans who previously had held key positions in the Zimbabwean government and other important sectors of the economy, left the country (Chimboza, 2012). Their reasons for leaving are attributed to negativity, fear, disillusionment and pessimism (towards the new independent Zimbabwe). The effect of this exodus was a national shortage of specialist skills in sectors such as education, health and engineering.

1991–1997: Economic Structural Adjustment Programme (ESAP), natural disasters and global recession

Poor economic performance and external pressure by international funders prompted the government to implement the Economic Structural Adjustment Programme (ESAP) in 1991. ESAP was introduced in the hope that it would encourage growth and employment, improve access to foreign exchange, and reduce deficit. However, the expected results of ESAP did not materialise, as was the case in most African countries that adopted ESAP. Instead, economic growth stifled, employment contracted, many firms closed, and social services deteriorated. Inflation rose and real wages decreased. Wages and salaries as a percentage of GDP fell from 57% in the 1980s to only 45% by 1995 (Mumbengegwi & Mabugu, 2001, cited in Chimboza, 2012). Most vulnerable groups became restless as standards of living deteriorated. Thousands of educated and skilled personnel left the country.

Besides putting up with the worst of ESAP, which was seen as an almost unqualified failure, the country experienced disastrous droughts in 1992 and 1995. A global recession in 1991/92 reduced raw material prices and export demand and trade. In addition, the country's biggest trading partner, South Africa, cancelled its trade agreement (Brett & Winter, 2003). In 1997, the government made a huge pay out to war veterans in capitulation to their demands for compensation in recognition of their sacrificial role for the country's independence (Brett & Winter, 2003). This payment was a political move that further weakened an ailing economy. Approaching the turn of the century, the country was thus on its way to bankruptcy.

1998–2008: Land reforms, sanctions, international isolation and hyperinflation

During the period 1998–2008, Zimbabwe went through the worst socio-economic challenges ever recorded in the country's history. The government engaged in land reforms that led to the collapse of commercial farming which was the backbone of the economy since memorial time. In 2000, disgruntled war veterans invaded white owned commercial farms. Land grabbing effectively crippled Zimbabwe's commercial industry, once dominated by 4,500 mainly white farmers and which, in the past, had constituted some 20% of the country's GDP and 40% of its export earnings (Besada & Moyo, 2008). Most land grabbers did not have the experience and infrastructure to pursue commercial farming. As a result, agricultural output dropped by 51% (UNESCO, 2015). Not only was the agricultural base of the economy destroyed, at least in the short to medium term, but also set Zimbabwe on what has become a protracted international conflict with the Western countries. What followed in 2002 were devastating economic sanctions on Zimbabwe from some Western countries and the cessation of funding from the World Bank and the International Monetary Fund (IMF). By 2003, most donors had suspended their operations in Zimbabwe. Expenditure on R&D was recorded as one of the lowest in the world. By May 2008, the official annual inflation rate had reached 1 million percent, the highest in the world (Besada & Moyo, 2008). During this period, 90% of the population was unemployed and 80% of Zimbabweans were living in poverty. This period saw the mass exodus of skilled professionals and intellectuals to neighbouring countries. Several universities were forced to close due to shortages of staff, food, water and continued student riots.

2009–2016: Government of National Unity, years of recovery and slow economic growth

Hyperinflation was brought into check in 2009 when the Government of National Unity (GNU) was put in power. This government consisted of the country's ruling political party and the main opposition party. Between 2009 and 2013, the new government embarked on some economic policies that created a more stable macro-economic environment. Once stabilized, the economy grew by 6% in 2009 and foreign direct investment increased slightly (UNESCO, 2014). It is important to note that the Zimbabwean economy in this period still remained fragile, plagued by high external debt, degraded infrastructure and an uncertain policy environment ADB-OECD-UNDP, 2014).

In sum, developments across these four periods can provide plausible explanations for patterns and trends that emerge from application of the classification of authorship types to agricultural and health research in Zimbabwe. The next section discusses the development of the classification framework and the relevant data sources and methods that were used.

Data sources, methods and classification

Compiling a database for this study was a creative endeavour, as it represents one of the few known attempts of integrating (at the record-level) different sets of articles from different data sources for an African country. The three sources were Scopus, the Web of Science (WoS), and the National Research Database of Zimbabwe (NRDZ) at the Research Council of Zimbabwe. The relevant period was from 1980 to 2016. Only articles were included as these reflect original research. The WoS articles were extracted from the database system of raw WoS article data at the Centre for Research on Evaluation, Science and Technology (CREST) at Stellenbosch University, South Africa. The Scopus articles were downloaded from the online system that the same university subscribes to. For both Scopus and the WoS, an article had to have at least one Zimbabwean author address in order to be included. The Scopus and WoS data was first unified to create a single set of non-duplicate data records. In the case of the NRDZ, all available article details were downloaded from the online database and checked against the Scopus-WoS dataset. Non-matching articles were individually scrutinised. These were subjected to full-text searches to identify articles published in predatory or potentially predatory journals. The Beall's list of predatory journals and publishers was used for this purpose, combined with other criteria (i.e. journals with questionable metrics and journals that ceased to exist after the first issues). In addition, NRDZ articles which were wrongly assigned to the article category and those without a Zimbabwean author address were also excluded.

Following multiple rounds of data cleaning and consistency checks, the final dataset (in Microsoft Access) comprised 11,056 articles during the period 1980–2016. Scopus contributed 2,759 (25%) unique articles to the final dataset, followed by WoS with 1,960 (18%) unique articles and the NRDZ with 277 (3%) unique articles each. However, in the fields of health and agricultural sciences (the focus of this study), only 29 and 3 articles in the NRDZ were from those fields respectively. These 32 NRDZ articles were published in 16 journals. Of these, 15 had been indexed in either Scopus or WoS at a particular point in time. The 32 missing articles, for various reasons, were lacking in Scopus and WoS.

A broad field classification framework for journals was developed, as informed by the framework used in a previous study (Boshoff, 2010), based on the WoS subject categories. Relevant journals from Scopus were assigned to the same set of WoS subject categories, by using information from the journal titles and field categories of Scopus journals. Any article (or rather journal) could be placed into more than one subject category. In turn, the subject categories were classified into six broad fields: agricultural sciences, engineering and technologies, health sciences, humanities, natural sciences, and social sciences. Two of the

broad fields (agricultural sciences [2094 articles] and health sciences [4382 articles]) constitute the focus of the current study.

A classification of authorship types (Table 1) was developed by creating and cross-tabulating four variables in the final article dataset. In other words, each article in the dataset was classified in terms of the following four variables:

- Number of article authors (coded as '1' and '≥2')
- Number of international countries authoring the article (coded as '0' and '1')
- Number of national organisations authoring the article (coded as '1' and '≥2')
- Type of national organisation, which could be either an international national organisation (INO) or a true national organisation (TNO).

Table 1. Classification of types of authorship.

				Number of international countries	
				0	**≥1**
Number of article authors	**1**	**Number of national organisations**	**1**	Single-authored articles • Type 1: TNO only • Type 2: INO only	Single-authored articles • Type 6: TNO only • Type 7: INO only
			≥2	Single-authored articles • Type 3: TNO only • Type 4: INO only • Type 5: TNO & INO	Single-authored articles • Type 8: TNO only • Type 9: INO only • Type 10: TNO & INO
	≥2		**1**	Nationally co-authored articles • Type 11: TNO only • Type 12: INO only	Internationally co-authored articles • Type 16: TNO only • Type 17: INO only
			≥2	Nationally co-authored articles • Type 13: TNO only • Type 14: INO only • Type 15: TNO & INO	Internationally co-authored articles • Type 18: TNO only • Type 19: INO only • Type 20: TNO & INO

Creating the first two variables was relatively straightforward, as it entailed determining the number of authors per article, and establishing whether an article listed at least one other country (apart from Zimbabwe) in the author address. The number of international countries per article reflected unique country names. Creating the third and fourth variables was more complicated. A non-duplicate list of standardised names of Zimbabwean author organisations first had to be compiled, whereafter those names were grouped into 13 sectors. Table 2 shows the classification of sectors, which each of the 556 unique Zimbabwean organisation names was placed into. Next, all Zimbabwean organisations were categorised as either an INO or TNO. All organisations belonging to any of the four sectors in the last column of Table 2 were classified as INOs and all organisations from the eight sectors in the first column as TNOs.

Table 2. Classification of Zimbabwean sectors and the alignment with TNOs and INOs.

Eight sectors involving TNOs	Four sectors involving INOs
• Zimbabwean university (e.g. National University of Science and Technology)	• Intergovernmental organisation (e.g. Food and Agriculture Organization – Zimbabwe)
• Zimbabwean national and local government (e.g. Department of Veterinary Services, Tsetse and Trypanosomiasis Control Branch)	• International research organisation/network or global or regional research partnership that operates in Zimbabwe (e.g. International Crops Research Institute for the Semi-Arid Tropics)
• Zimbabwean NGO/community based organisation/faith based organisation (e.g. Iluba Elimnyama Theatre Works)	• International NGO/philanthropic organisation/foundation/think-tank that operates in Zimbabwe (e.g. Elizabeth Glaser Pediatric AIDS Foundation)
• Zimbabwean industry/business/company/firm (e.g. National Foods Limited)	• International business/company/firm that operates in Zimbabwe (e.g. Deloitte and Touche Zimbabwe)
• Zimbabwean private school/college/university (e.g. Christian Brothers College)	Note: There is also a 13th sector, called 'other', which includes, for instance, unknown street and postal addresses.
• Zimbabwean private clinic/hospital (e.g. Royal Women's Clinic)	
• Zimbabwean mission/faith-based hospital (e.g. Sanyati Baptist Hospital)	
• Zimbabwean union/association (e.g. Zimbabwe Psychological Association)	

Given the above explanation of classification procedures, interpretation of the different types of authorship in Table 1 should now be effortless. For instance, type 1 represents a single-authored article, produced by a single Zimbabwean organisation without any international co-authorship, and where that organisation could be any of the eight sectors in the first column of Table 2. Type 20, on the other hand, represents an internationally co-authored article because of at least one international country address. At the same time, the article also involves national co-authorship, given the listing of two or more authors together with two or more Zimbabwean organisational addresses. At least one of the Zimbabwean organisations is a TNO (belonging to any of the eight sectors in the first column of Table 2) and at least one other an INO (belonging to any of the four sectors in the last column of Table 2).

Results

Table 3 was constructed by applying the classification of authorship types to Zimbabwe's article output in agricultural research in the four socio-economic periods. The periods are composed of different numbers of years. The first and third period each comprises 11 years, whereas the second period is composed of seven years and the last of eight years. For that reason percentage breakdowns should be considered when comparing the results for the socio-economic periods, given that percentages are a form of standardisation and hence unaffected by the differences in years. Contributions of 10% or more to the total agricultural research output in a particular period are highlighted.

Table 3. Types of authorship in agricultural research, by socio-economic period.

Types	Socio-economic periods							
	1980–1990		1991–1997		1998–2008		2009–2016	
	Count	%	Count	%	Count	%	Count	%
Type 1	106	34%	62	14%	37	5%	9	2%
Type 2*	5	2%	11	2%	4	1%	1	<1%
Type 3	0	0%	1	<1%	1	<1%	0	0%
Type 4*	0	0%	0	0%	0	0%	0	0%
Type 5*	0	0%	0	0%	0	0%	0	0%
Type 6#	2	1%	4	1%	3	<1%	2	<1%
Type 7*#	1	<1%	0	0%	1	<1%	1	<1%
Type 8#	0	0%	0	0%	0	0%	0	0%
Type 9*#	0	0%	0	0%	0	0%	0	0%
Type 10*#	0	0%	0	0%	0	0%	0	0%
Type 11	88	28%	142	32%	75	10%	2	<1%
Type 12*	6	2%	14	3%	4	1%	1	<1%
Type 13	19	6%	32	7%	93	12%	125	22%
Type 14*	0	0%	2	<1%	2	<1%	17	3%
Type 15*	4	1%	2	<1%	12	2%	25	4%
Type 16#	60	19%	119	26%	284	37%	130	23%
Type 17*#	7	2%	21	5%	125	16%	76	13%
Type 18#	9	3%	35	8%	102	13%	92	16%
Type 19*#	0	0%	1	<1%	11	1%	42	7%
Type 20*#	2	1%	4	1%	18	2%	40	7%
Total	**309**	**100%**	**450**	**100%**	**772**	**100%**	**563**	**100%**

* Article involving an INO # Article involving international co-authorship (IC)

According to Table 3, in the first period (1980–1990), articles in agricultural research were mainly produced by a single Zimbabwean organisation (classified as a TNO), involving either one author (34%, type 1) or more authors (28%, type 11). The percentage contribution of type 1 to the total article output in agricultural research decreased systematically and significantly between adjacent periods. On the other hand, the percentage contribution of type 11 first increased before taking a sharp decline following the second period. Type 16, which indicates co-authorship between a single TNO and one or more international partners, had a consistent presence in all four periods, especially in 1998–2008 when it peaked at 37%. Together with type 13 (national co-authorship between two or more TNOs), type 16 dominated the last period (23% and 22%), with types 17 (13%) and 18 (16%) also having a notable presence in that period. Type 17, or co-authorship between a single INO and an international country, is associated with the largest contribution by INOs to the total article output in agriculture. This is particularly true for the period 1998–2008 (16%).

The reporting structure of Table 4 replicates that of Table 3, the difference being that the focus is on health research. Type 1 (36%) and type 11 (39%), collectively, were responsible for three-quarters of articles in health research in the period 1980–1990 but for less than 3% in the period 2009–2016. Types 13 and 16 are prominent across all four periods but display different trends. Type 13 decreased after the second period whereas type 16 recorded consistent growth. Type 16 (33%) and type 18 (37%) dominated the last period, with type 13 (12%) and type 20 (11%) also enjoying visibility in that period. The 6% for type 20, together with the 6% for type 17 (2009-2016) and the 8% for type 17 in the preceding period (1998–2008), constitute the three largest contributions by INOs.

Table 4. Types of authorship in health research, by socio-economic period.

Types	Socio-economic periods							
	1980–1990		1991–1997		1998–2008		2009–2016	
	Count	%	Count	%	Count	%	Count	%
Type 1	335	36%	128	14%	87	7%	26	2%
Type 2*	0	0%	7	1%	8	1%	0	0%
Type 3	1	<1%	0	0%	0	0%	5	<1%
Type 4*	0	0%	0	0%	0	0%	0	0%
Type 5*	0	0%	0	0%	0	0%	0	0%
Type 6#	4	0%	4	<1%	10	1%	9	1%
Type 7*#	0	0%	0	0%	1	<1%	1	<1%
Type 8#	0	0%	0	0%	1	<1%	0	0%
Type 9*#	0	0%	0	0%	0	0%	0	0%
Type 10*#	0	0%	0	0%	0	0%	0	0%
Type 11	367	39%	226	25%	69	6%	1	<1%
Type 12*	2	<1%	1	<1%	3	<1%	0	0%
Type 13	106	11%	186	21%	216	18%	163	12%
Type 14*	0	0%	1	0%	1	<1%	2	<1%
Type 15*	2	<1%	6	1%	9	1%	12	1%
Type 16#	92	10%	237	26%	414	34%	429	33%
Type 17*#	5	1%	24	3%	100	8%	85	6%
Type 18#	21	2%	76	8%	283	23%	481	37%
Type 19*#	0	0%	2	<1%	5	<1%	17	1%
Type 20*#	1	<1%	0	0%	25	2%	85	6%
Total	**936**	**100%**	**898**	**100%**	**1232**	**100%**	**1316**	**100%**

* Article involving an INO # Article involving international co-authorship (IC)

Figure 1 compares the overall contribution of INOs to agricultural research and health research in Zimbabwe, by socio-economic period. In other words, it shows the combined contributions of authorship types that are marked with an asterisk (*) in Tables 3 and 4. Both fields show consistent growth in terms of the participation of INOs. In the period 2009–2016, more than one-third (36%) of articles in agriculture involved an INO. For health, the figure for that period was 15%. This represents a significant increase over the 1% recorded for the early years of independence (1980–1990).

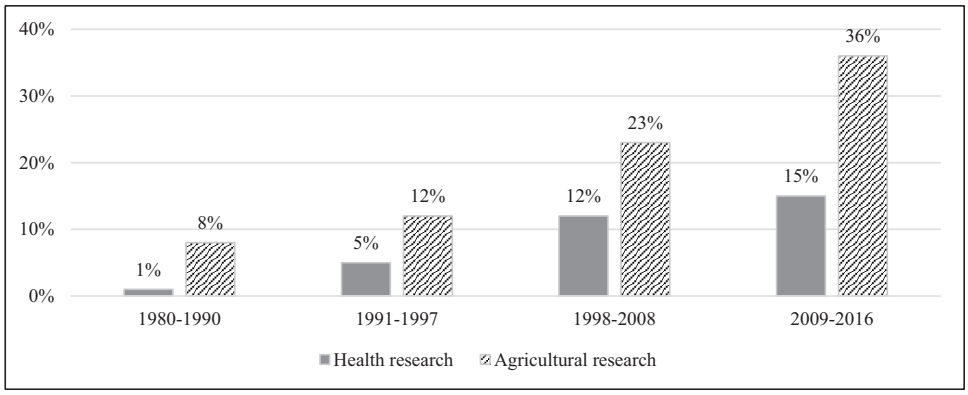


Figure 1: Percentage contribution of INOs to agricultural research and health research in Zimbabwe, by socio-economic period.

Table 5 lists the 8 most frequently occurring INOs in agricultural research and in health research in the last two periods combined. The type of INO appears in brackets after the organisation name (see the table footnote). According to Table 5, the International Maize and Wheat Improvement Center (CIMMYT), which leads among INOs in agriculture in Zimbabwe (41%), produces both agricultural and health research. This is because of the multiple subject classification of journals in which it publishes, like *Food and Nutrition Bulletin* and *Food Policy*. The CIMMYT, with its headquarters in Mexico, is hosted by a number of developing countries.

The most frequently counted INO in health is the World Health Organisation (WHO), accounting for 41% of all article output by INOs in the two relevant periods. The United Nations Children's Fund (UNICEF), Population Services International and the Centers for Disease Control and Prevention all follow in second place with 6%.

Table 5. Top 8 INOs operating in agricultural research and health research in Zimbabwe, 1998–2016.

Agricultural research			Health research		
Organisation	**Count**	**%**	**Organisation**	**Count**	**%**
International Maize and Wheat Improvement Center (IRO)	156	41%	World Health Organisation (IGO)	145	41%
International Crops Research Institute for the Semi-Arid Tropics (IRO)	92	24%	United Nations Children's Fund (IGO)	23	6%
International Centre for Tropical Agriculture (IRO)	36	9%	Population Services International (ING)	21	6%
World Agroforestry Centre (IRO)	23	6%	Centers for Disease Control and Prevention (IRO)	21	6%
UF/USAID/SADC Heartwater Research Project (IRO)	14	4%	International Maize and Wheat Improvement Center (IRO)	17	5%
Center for International Forestry Research (IRO)	13	3%	UF/USAID/SADC Heartwater Research Project (IRO)	16	4%
Cirad - Zimbabwe (IRO)	12	3%	Elizabeth Glaser Pediatric AIDS Foundation (ING)	13	3%
Seed Co - Zimbabwe (INB)	9	2%	Letten Foundation Research Center (IRO)	12	3%

IGO = Intergovernmental organisation; IRO = International research organisation/network or global or regional research partnership that operates in Zimbabwe; INB = International business/company/firm that operates in Zimbabwe; ING = International NGO/philanthropic organisation/foundation/think-tank that operates in Zimbabwe.

Finally, Figure 2 shows the overall international participation in agricultural research and health research, where international participation comprises three components: (1) articles involving international co-authorship (IC) only, (2) articles involving both IC and an INO, and (3) articles involving an INO only.

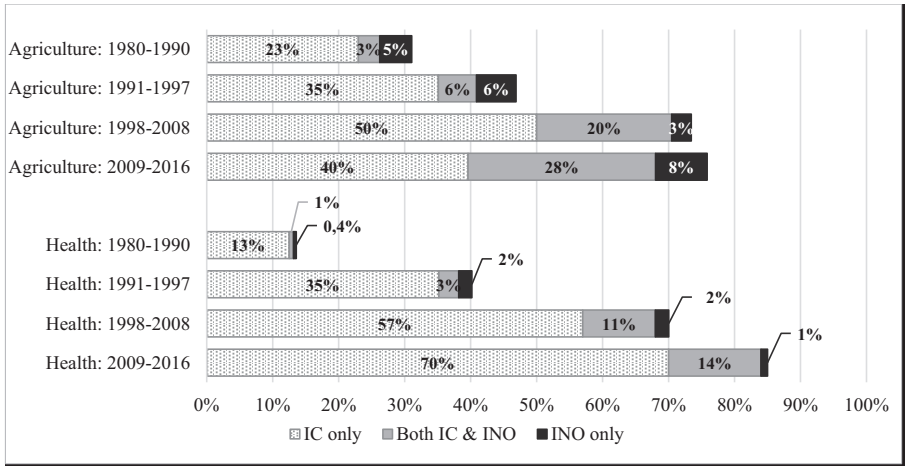


Figure 2. Percentage breakdown of forms of international participation in agricultural research and health research in Zimbabwe, by socio-economic period.

In both agriculture and health, but especially in health, participation of INOs without international co-authorship also being simultaneously present, was found to be the exception rather than the rule. In the period 2009–2016, respectively 8% and 1% of research articles in agriculture and health were produced by INOs without the presence of international co-authorship. A first observation therefore is that participation of INOs in Zimbabwe's agricultural and health research seldom happens without international co-authorship also being present simultaneously. This raises a question about the research link between international co-authorship and INOs. A second observation is that the contribution of INOs to international co-authorship significantly increased with time. For instance, in agriculture, 26% of agricultural research output in the period 1980–1990 involved international co-authorship, of which 3% also included INO participation. By 2009–2016, 68% of the total output involved international co-authorship, of which 28% listed participation by an INO.

Discussion

This study contributed in three ways to the understanding of research production in a developing African country like Zimbabwe. First, it expanded the notion of international participation by including INOs as another form of such participation. Second, it provided a new set of authorship categories (20 types) through which to view research production in the developing context and, third, it provided empirical support to the view that bibliometric analyses should be conducted and interpreted in context. Based on the overview of four socio-economic periods in Zimbabwe, preliminary ideas might have taken shape about how the nature of changes in the country's socio-economic environment would affect the results of the study. For instance, it might have been expected that international participation would decrease between 1998 and 2008 because, during this period, the country faced international isolation while funding institutions like the World Bank and the IMF ceased their funding obligations, and, worse still, economic sanctions were levelled to the country. However, the results showed that international participation in both agricultural and health research consistently and significantly increased, even while the country was in turmoil and internationally isolated. Two factors might have been at play here, accounting for this anomaly – the internal workings of science (e.g. networking practices in agricultural and health research at individual and organisational level) and a global urge to develop research partnerships across international divides to address grand societal challenges (e.g. food security and HIV/Aids).

Application of the framework of authorship types to Zimbabwe's research in agriculture and health showed that, during the early years of independence, research in the two fields was predominantly produced by one or more authors from a single national institution. However, as time passed, research became largely dominated by international participation. One explanation for this trend could be that, at the time of the economic meltdown during the period 1998–2008, most skilled professionals and researchers had left the country. Mouton et al. (2008, 250) remark that the human capital base in Zimbabwe had "been eroded to the point where effective research and teaching was barely possible". Hence, the few remaining researchers who wanted to pursue a research agenda, as well as the newly trained researchers when they returned from their international scholarships, could only do so through international co-authorship, which included participation by INOs. The other explanation could be that during the economic crisis, researchers in the country became increasingly dependent on international authors for resources, and thus diligently pursued all possible avenues of international networking.

Participation by INOs, without any co-current international co-authorship, was found to be minimal. For instance, only about 8% of articles in agriculture in 2009–2016 listed a Zimbabwean INO in the author address without also mentioning an international co-author. Moreover, an increase in international co-authorship was found to coincide with an increase in INOs. The link between international co-authorship and INOs thus requires further investigating. Various hypotheses should to be explored in this regard. For instance, it could be a matter of multiple affiliations, where some international co-authors of Zimbabwean articles – reflecting an international organisation (IO) – are associated with an INO in Zimbabwe. For this reason, the article data for the period 2009–2016 was converted into a dataset of Zimbabwean authors. Table 5 shows the number of unique Zimbabwean authors in each field, and the breakdown across different combinations of TNOs, INOs and IOs.

Table 5: Breakdown of author affiliations in agricultural research and health research in Zimbabwe, 2009–2016.

	Total number of Zimbabwean authors	Authors broken down in 6 mutually exclusive categories					
		TNO only	INO only	TNO & INO	TNO & IO	INO & IO	TNO, INO & IO
Agricultural research	655	423	61	19	113	32	15
Health research	1393	959	129	11	222	40	24

Based on Table 5, it is observed that, between 2009 and 2016, a total of 423 Zimbabwean authors in agriculture were affiliated with TNOs only. The remaining 35% of authors in agriculture therefore had some international link in one way or the other (i.e. reporting either an IO or an INO as their author address in at least one article in the relevant period). Specifically, 160 (24%) of Zimbabwean authors in agriculture reported an IO affiliation in the relevant period. Moreover, 47 (29%) of the authors affiliated with an IO also had an INO address. In the case of health, 64 (22%) of the 286 Zimbabwean authors with at least one IO affiliation also had an INO address.

In a follow-up study, based on articles as the unit of analysis, specific attention would need to be paid to the patterns of co-authorship between specific INOs and specific countries, and the participation of other Zimbabwean organisations (TNOs) in the co-authorships. Attention should also be paid to topic modelling, in order to determine clusters of research topics and whether the clusters differ according to the involvement of TNOs, INOs and IOs in the research.

Finally, although the finding of increased international collaboration in Africa's research is well established (e.g. Adams et al., 2014; Mouton & Blanckenberg, 2018), the observation that it

coincides with participation by INOs is not. The current study thus illustrates how small tailor-made bibliometric datasets, developed for African countries with small science systems, could in fact produce new insights and frameworks to inform future studies that investigate research landscapes in the developing world.

References

Adams, J., Gurbey, K. & Hook, D. (2014). International collaboration clusters in Africa. *Scientometrics*, 98, 547-556.

ADB-OECD-UNDP (2014). *African Economic Outlook 2014. Global Value Chains and Africa's Industrialisation*. Report by the African Development Bank, Organisation for Economic Co-operation and Development and United Nations Development Programme Retrieved January 2 2019 from www.africaneconomicoutlook.org/en.

AU-NEPAD (2010). *African Innovation Outlook 2010*. Pretoria: African Union-New Partnership for Africa's Development.

Besada, H. & Moyo, N. (2008). *Zimbabwe in Crisis: Mugabe's Policies and Failures*. Working paper No.38 Retrieved November 22, 2018 from: www.cigionline.org.

Boshoff, N. (2009). Neo-colonialism and research collaboration in Central Africa. *Scientometrics*, 81, 413-434.

Boshoff, N. (2010). South-South research collaboration of countries in the Southern African Development Community (SADC). *Scientometrics, 84*, 481-503.

Brett, T. & Winter, S. (2003). *Origins of the Zimbabwe Crisis*. Retrieved January 17 2019 from: https://hsf.org.za.

Chimboza, A. (2012). *From Brain Drain to Brain Gain: Addressing Human Capital Needs for Post Crisis Zimbabwe's Capacity Building*. Master of Philosophy Theses. 11. Retrieved September 22, 2017 from: http://repository.upenn.edu/od_theses_mp/11.

Katz, J.S. & Martin, B.R. (1997). What is research collaboration? *Research Policy, 26*, 1–18.

Kozma, C., Medina, C.C. & Costas, R. (2018). Research funding landscapes in Africa. In C. Beaudry., J, Mouton & H. Prozesky, (Eds). *The Next Generation of Scientists in Africa* (pp.26-42). African Minds: South Africa.

Landini, F., Malerba, F. & Mavilia, R. (2015). The structure and dynamics of networks of scientific collaborations in Northern Africa. *Scientometrics*, 105, 1787-1807.

Mêgnigbêto, E. (2013). *Scientific research in West Africa: A global view. (2001-2010).* Retrieved June 29, 2017 from: http://eustachem.ilemi.net/IMG/pdf/megnigbeto-issi-2013.pdf.

Mouton, J., Boshoff, N., Waal de. L., Esau, S., Imbayarwo, B., Ritter, M. & Niekerk. D. (2008). The state of public science in the SADC region. In P. Kotecha, (ed). *Towards a common future: Higher education in the SADC region. Research findings from four SARUA studies* Johannesburg: Southern African Regional Universities Association.

Mouton, J. & Blanckenberg, J. (2018). African science: A bibliometric analysis. In C. Beaudry., J, Mouton & H. Prozesky, (Eds). *The Next Generation of Scientists in Africa* (pp.26-42). African Minds: South Africa.

Owusu-Nimo, F. & Boshoff, N. (2017). Research collaboration in Ghana: Patterns, motives and roles. *Scientometrics*, 110(3), 1099-1121.

Pouris, A. & Ho, Y-S. (2014). Research emphasis and collaboration in Africa. *Scientometrics*, 98, 2169-2184.

Scientific Industrial Research and Development Centre. (2003). An analysis of the cause and effect of brain drain in Zimbabwe. Harare: SIRDC.

Sugimoto, C.R. & Lariviere, V. (2018). *Measuring Research: what everyone needs to know*. Oxford: University Press.

UNESCO (2014). *Mapping research and innovation in the Republic of Zimbabwe*. *Vol 2*. Retrieved June 27, 2017 from: file:///C:/Users/19655118/Downloads/GO-Spin-%20Zimbabwe%20report%20(1).pdf.

UNESCO (2015). *UNESCO Science report. Towards 2030*. Retrieved October 23, 2018 from: https://en.unesco.org/unesco_science_report.

Zogbo, J.R. (2003). *The Impact of the Economic Structural Adjustment Programme on Education In Zimbabwe*. Retrieved January 27, 2019 from: https://opendocs.ids.ac.uk.

Using altmetrics to study social movements and cognitive bridges in the communication of science in the social media: The case of the anti-vaccination movement on Twitter

Francois van Schalkwyk,[1] Jonathan Dudek[2] and Rodrigo Costas[1,2]

[1] *fbvschalkwyk@sun.ac.za*
Centre for Research on Evaluation, Science and Technology (CREST), Stellenbosch University, Private Bag X1 Matieland, 7602, Stellenbosch (South Africa)

[2] *j.dudek@cwts.leidenuniv.nl, rcostas@cwts.leidenuniv.nl*
Centre for Science and Technology Studies (CWTS), Leiden University, Wassenaarseweg 62A, Leiden, 2333AL (The Netherlands)

Abstract

This paper presents an analysis of the anti-vaccination movement's interactions with open access scholarly papers on Twitter. By using tweeter coupling analysis combined with data on the stance of Twitter accounts, the networks and sub-networks of tweeters around the debate of anti-vaccination and autism are plotted, analysed and discussed. The main findings indicate distinct communities of shared interests within the anti-vaccination movement, and that these communities are highly connected to one another within the anti-vaccination movement and relatively disconnected from a pro-science community within the same network. Communities with shared interests harbour more active accounts that are described as cognitive bridges within and between clusters in the network. These are possibly indicative of different types of intermediaries in the flow of scientific information in the network. This paper represents an attempt at creating an empirical understanding of the potential of altmetrics to study science–society interactions; more specifically, the potentials arising from increased access for non-scientists to the research products produced in the formal communication of science.

Introduction

The increase in advocacy for transparency and accountability, operationalised as openness and access, stems in part from a degradation of trust in public institutions (Castells, 2009, 2017; Ortiz-Ospina & Roser, 2016), including those institutions tasked with conducting scientific research. The demands for accountability through greater transparency, oversight and measurement of public institutions are buttressed by claims of beneficial returns to society (Weingart, 2012). Open science is, from such a point of view, seen as a necessary evolution towards improvements in the efficiency, quality and contribution of science to society (Jasanoff, 2006; Leonelli, Spichtinger, & Prainsack, 2015).

Advances in technology have radically transformed the interconnectedness of society, with new modes of communication emerging. The effects of such changes for the communication of science become more relevant in the light of open science, which increases access to the formal communications of science by the public.

For science, the emergence of information and communication technologies in society has seen a range of impacts on its communication. The increase in the number of open access journals (Archambault, et al., 2014; Piwowar, et al., 2017), the ability of a greater number of publics to access journal articles via the internet, new modes of knowledge exchange between science and its publics (Dickel & Franzen, 2016), and the presence of journal articles on social media platforms such as Twitter (Haustein, Costas, & Lariviere, 2015; Thelwall, Tsou, Weingart, Holmberg, & Haustein, 2013) while the use of social media by scientists remains relatively low (Bar-Ilan, et al., 2012; Van Noorden, 2015), suggest that journal articles are indeed incorporated into the communications of publics on social media platforms (Haustein & Costas, 2015).

Communication networks are the patterns of contact and exchange created by flows of signals and information between communicators through time and space (Monge & Contractor, 2003).

As actors interact in social media spaces such as Twitter, they form connections that emerge into complex social network structures that reflect patterns of information flows.

From those network structures, clusters of proximal social actors emerge. Clusters in communication networks may emerge because of shared aspiration, purpose or identity. The clusters may constitute communities or social movements, collectives of social actors acting self-consciously to effect change of some kind (Stalder, 2006). Such social movements are no longer bound by place and are more readily able to function in real time on a global scale. For better or for worse, their globalisation has increased their effectiveness. Some sociologists therefore remain ambivalent about the social potential of social movements in the network society (Miller, 2017; Stalder, 2006).

Open science makes possible increased public access to science and new communication technologies provide the space for interactions between scientists and diverse publics, including social movements. This development raises concerns around how non-scientific, ideologically motivated publics self-organised into social movements access and distribute scientific information from journal articles as part of their communication strategies. One example of such a social movement attentive to science is the anti-vaccination movement. As far back as in 2005, researchers were aware of the potential harm that could result from the use of science as a source of health information by ideologically motivated social movements (Zimmerman, Wolfe, & Fox, 2005). Others (Bennato, 2017; Bean, 2011; DiResta & Lotan, 2015; Kata, 2012; Leask, 2015) elaborate on this concern, arguing that the effects of activist groups should not be underestimated. Through the social media in general, and Twitter in particular, anti-vaccination promotors could yield considerable influence over the general public, particularly among those faced with decisions about whether to vaccinate their children.

The media have traditionally been the primary interface between science and the public (Weingart, 2011), and it is the science journalist who has kept the public informed on the latest developments from the world of science (Schäfer, 2017). There is, however, a simultaneous decline in science journalism (Scheufele, 2013; Schäfer, 2017), an increase in the scramble for attention among a variety of stakeholders (Weingart & Guenther, 2016; Williams 2018), and the emergence of social media as a new, informal, interpersonal channel of communication between scientists and the public (Southwell, 2017). Bucchi (2018) describes a 'crisis of mediators' in which new scientific research is increasingly fed in real time into the public domain without being filtered by communication professionals. As the knowledge products of science, including journal articles, become more accessible, attentive non-scientist publics are presented with a pluralistic universe of information providers and must choose whom to trust as authoritative sources of 'scientific' information (Blöbaum, 2016).

In online communication networks, intermediaries assume the form of centrally located actors (Feng, 2016) who are in a position to exercise power as influence (Muller, 2017; Williams, 2018) and who, by virtue of their position in or between networks, are able to make new connections (Burt, 2001; Feng 2016) or, more importantly, challenge the dominant programme of a specific network (Castells, 2009).

As a socially constructed space, the organisation of actors and their relationships in communication networks in an age of information (Castells 2009) is key to understanding the delivery, reception, use and impact of science. We expect social movements to play a strategic role in the amplification of messaging extracted from openly accessible products of science in their online communication networks. Individual actors may mediate in new ways (Landrum, 2017; Scheufele, 2014) communications between science and a social movement such as the anti-vaccination movement. Hence, this paper seeks to establish whether the anti-vaccination movement is present in a network predicated on the flow of scientific information in the social media, and whether there are possible intermediaries in that network.

Methodology

By means of a case study, the research presented here is centred around social media activities of the anti-vaccination movement, more specifically the interactions on Twitter between members of that movement and open access journal articles on the topic of vaccination and autism. This selection constitutes a viable object of research, as it allows for the observation of a non-scientific community with an explicit interest in and interaction with scientific findings.

All communication under study here is focused on open access journal articles. This decision not only follows the thematic prerequisite of investigating the potential risks of open science in the network society, it also ensures that any publication shared in social networks is not limited regarding the accessibility of its content. For this study, a set of 113 open access journal articles on the topic of vaccination and autism was used. This set was taken from a study by Van Schalkwyk (2019) on the potentials of open science.

In order to reconstruct networks from those publications, all Twitter accounts that mentioned any of the 113 articles were identified, as well as the number of tweets per Twitter account. This information was retrieved from data provided by Altmetric.com, with the latest date of tweeting events being 31 October 2017. This returned a total of 12,207 unique Twitter accounts and a total of 21,490 tweets.

Following the data collection from the Altmetric.com database, a network was created on the basis of both Twitter accounts and their tweets mentioning an open access journal article from the predefined set.

To create this network, the approach of bibliographic coupling (Kessler, 1963; Zhao & Strotmann, 2008) was adapted for the social media and its affordances. Multiple products of science in the form of open access journal articles (linked objects) were used to create instances where two Twitter accounts (tweeters or users) tweet a link (URL) to the same set of open access journal articles. Each pair of Twitter accounts linking to a journal article in this manner is a couple, hence the network is described as a 'tweeter coupling network' (Costas et al., 2017). Figure 1 illustrates this concept.

Coupling based on journal articles indicates a type of coupling in which a cognitive link or bridge is established between two Twitter accounts. A *cognitive bridge* in this context is a connection that exists between two actors who share a common interest in an idea or set of ideas. It is not necessarily the case that there is cognitive alignment or harmony between bridged actors; rather, what bridges them is their shared interest in a particular idea or set of ideas. Cognitive coupling in the social media is distinct from, for example, social or semantic coupling which are premised on other links such as followers or hashtags respectively.

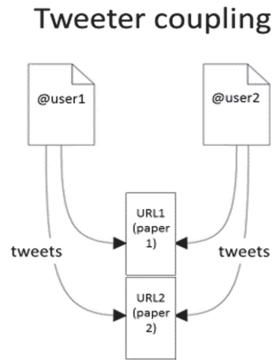


Figure 1. Coupling of tweeters over shared publications (Costas et al., 2017)

In the manner described above, a matrix of coupled Twitter accounts was created. The data consisting of Twitter account pairs and the numbers of their co-occurrences was further processed with the software package NodeXL Pro version 1.0.1.396 (Hansen, Shneiderman, & Smith, 2010). This software was used for all network-metric calculations and for generating the respective network graphs. It makes possible the plotting of a network graph showing the connections between those Twitter accounts that mentioned more than two open access journal articles from the sample.

Using a web crawler and manual analysis of Twitter accounts flagged by the crawler, Van Schalkwyk (2019) identified from 10,544 Twitter accounts, 658 anti-vaccination accounts that mention one of ten open access journal articles on the topic of vaccination and autism. From the verification process, several pro-science accounts were also identified. Anti-vaccination Twitter accounts were defined as those that regularly tweet or retweet content to persuade others of the dangers of vaccines, while pro-science accounts were those that tweet to defend the consensus position of science, that is, that vaccines are effective in combatting infectious disease and pose no material health risks to those who are vaccinated. The data on stance, while not comprehensive in the sense that it provided classifications for all tweeters in the tweeter coupling network for 113 articles, was added to NodeXL to provide additional information in the analysis of the network.

Findings

Figure 2 shows a tweeter coupling network in which each pair of Twitter accounts shares a mention to more than 2 of the 113 open access journal articles related to the topic of vaccination and autism. Red nodes show *anti-vaccination* accounts while green nodes show *pro-science* Twitter accounts; the stance of the black nodes is unknown.

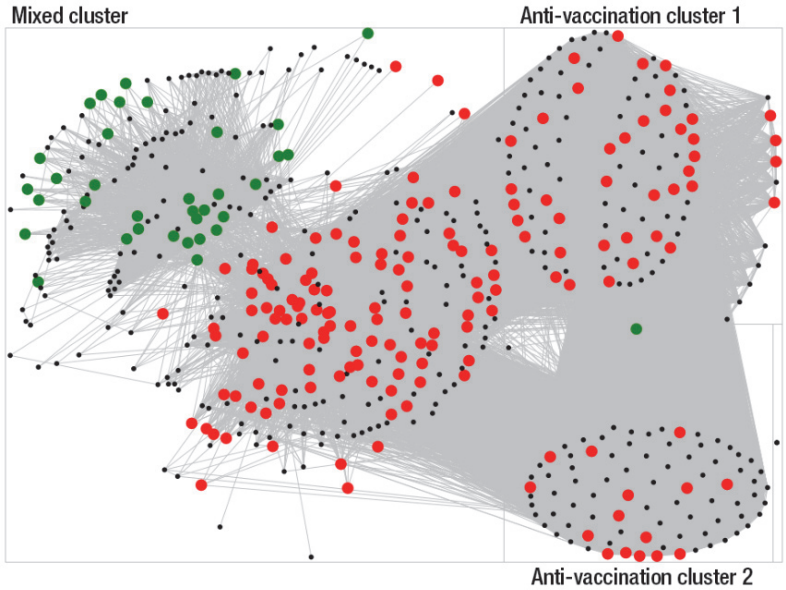


Figure 2: Tweeter coupling network (*N* = 12 207)

The graph reveals several noteworthy characteristics. The first and most obvious is that the openly accessible formal communications of science (i.e. open access journal articles) are being

mentioned on the social media platform Twitter and that it is possible to generate a tweeter coupling network based on the interactions between tweeters and these publications.
The second noteworthy characteristic of the graph is the clustering of Twitter accounts into three distinct groups. The two groups on the right-hand side of the graph are composed of anti-vaccination accounts. The third "mixed group" (on the left-hand side of the graph) consists of both pro-science and anti-vaccination accounts, although there is also a discernible division within this group according to stance: pro-science accounts are clustered in the top left while anti-vaccination accounts are clustered in the bottom right and closer to the two exclusively anti-vaccination groups.
Other notable characteristics of this graph include the fact that the clusters exhibit both high degrees of interconnectedness (between clusters) as well as high degrees of intra-connectedness (within clusters). The vertices between the three anti-vaccination clusters are demonstrably denser than the links between the anti-vaccination sub-group and the pro-science Twitter accounts in the same group.
Algorithms for calculating centrality extend the analysis beyond communities of shared interests to the positions of individual tweeters relative to others in the same network. This provides insight into who the most active coupled tweeters are in the sense that they display topical dominance in a network predicated on journal articles dealing with the topic of vaccination and autism. Table 1 shows those network nodes with the highest degree centrality scores, i.e. those nodes with greatest number of connections to other nodes in the network. All tweeters with the highest degree centrality scores are anti-vaccination. The account with the highest degree centrality is @itsmepanda1 (404). The pro-science Twitter account with the highest degree centrality score is @dkegel (140), followed by @doritmi (135).

Table 1. Top 10 accounts by degree centrality score in the tweeter coupling network

Account	Stance	Degree
@itsmepanda1	Anti-vaccination	404
@LaLaRueFrench75	Anti-vaccination	365
@debnantz	Anti-vaccination	344
@eTweeetz	Anti-vaccination	340
@Biegenzahn	Anti-vaccination	335
@libertylives277	Anti-vaccination	319
@EMcCra2	Unknown	318
@aspiritcan	Anti-vaccination	309
@SNCCLA	Anti-vaccination	304
@VaxChoiceVT	Anti-vaccination	303

Typically, the distribution of degree centrality in social networks follows a power law distribution (Kahle, 2011). Figure 3 shows that the distribution is typical of a social network – a high number of accounts with low centrality scores, followed by a rapid drop-off in the number of low degree centrality accounts, and a long tail of accounts with diminishing high degrees of centrality. What is atypical are the 'spikes' in mid-range degree centrality at scores in the following bands: 120-130, 150-160 and 220-230.
A closer examination of the bands in which these spikes occur reveals that all three bands consist of Twitter accounts located in the anti-vaccination clusters, primarily the two exclusively anti-vaccination clusters. The accounts in these bands link to an unusually high number of other accounts in the network. These anomalies in the mid-range degree centrality scores indicate that tweeters not only link with one another to different degrees but that there are an unusually high number of active tweeters when it comes to interacting with open access journal articles by linking to them in their tweets. In other words, there are what one might term anti-vaccination super-disseminators who exhibit topical dominance and who appear more frequently in the network than one would normally expect to be the case.

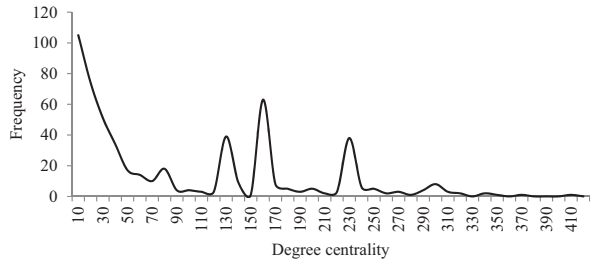


Figure 3. Distribution of degree centralities in the tweeter coupling network

Figures 4 and 5 show two different patterns of cognitive coupling in the anti-vaccination movement by focusing on two Twitter accounts, one with the highest degree centrality (Table 1) and one tweeter known to be highly active in the anti-vaccination movement more broadly (Van Schalkwyk, 2019). Figure 4 shows that the highly active anti-vaccination Twitter account only spans the three anti-vaccination clusters suggesting that this account as a cognitive bridge is limited to bridging *within* the anti-vaccination movement. Figure 5 shows how the anti-vaccination Twitter account with the highest degree of centrality in the network spans the three anti-vaccination clusters in the network, and spans from the anti-vaccination accounts to pro-science cluster in the network. This suggests that this account represents a cognitive bridge between sub-networks within an aligned group, as well as between non-aligned sub-networks.

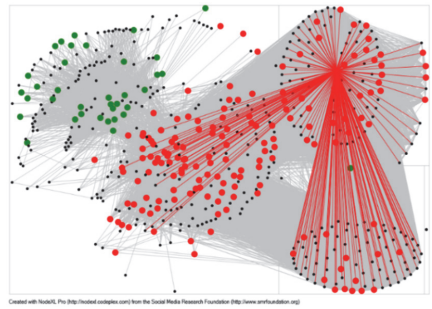


Figure 4. Example of a cognitive bridge in a tweeter coupling network: @LotusOak

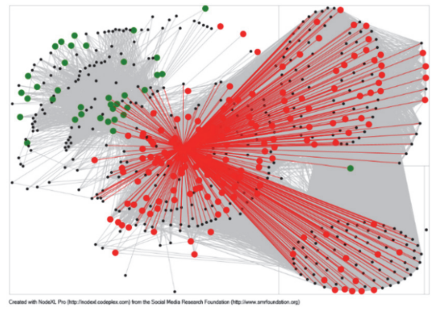


Figure 5. Example of a cognitive bridge in a tweeter coupling network: @itsmepanda1

Figure 6 shows only those tweeter couples that mentioned seven or more open access journal articles (seven being roughly the mid-way point between the minimum and maximum number of mentions). The graph confirms that most active tweeter couples are anti-vaccination; that there are tweeter couples that consist of accounts in different anti-vaccination sub-groups; and that the smaller group of pro-science couples are isolated from the anti-vaccination clusters with the exception of one couple consisting of a known pro-science tweeter and an anti-vaccination tweeter.

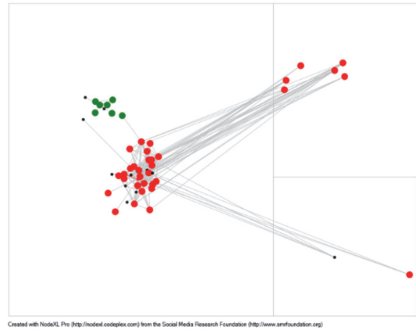


Figure 6. Coupling network where mentions are greater than 7 open access journal articles

Discussion

This paper set out to discover the composition of the anti-vaccination movement based on its interactions with relevant open access journal articles. The findings from a network analysis of those interactions indicate that rather than being a single community, there are distinct sub-groups that constitute the anti-vaccination movement. At the same time, there appear to be different types of centrally located actors in the communication network of the anti-vaccination movement, each ostensibly playing important roles in the dissemination of scientific information.

The two exclusively anti-vaccination clusters or groups are arranged into sub-networks that reveal a typical hub-and-spoke arrangement (Himelboim et al., 2017). These groups are centred around one or two Twitter accounts. However, unlike hub and spoke networks, the density of connections between the three anti-vaccination clusters is relatively high. This is attributable to high levels of cognitive coherence and could possibly indicate the highly selective exchange of information between clusters.

The mixed group consisting of both anti-vaccination and pro-science accounts shows a different type of network structure; one that can be described as polarised (Himelboim et al., 2017). In this part of the network, links between the pro-science and the nearest anti-vaccination cluster are less dense while in-cluster density remains high. Here, there are links between anti-vaccination and pro-science tweeters, indicating cognitive bridges between the two opposing groups. Although there are links between the pro-science and the anti-vaccination clusters, they are few compared to dense links between the anti-vaccination clusters. This suggests that Twitter accounts in the network are primarily exposed to purposively selected content that aligns with the ideology of the broader movement. Polarisation might be a consequence of this, leaving little room for more nuanced positions or constructive engagement (Tucker et al., 2018; Yeo et al., 2015).

Barberá (2015) challenges the widely held view that social media intensifies polarisation by positing that social media usage reduces political polarisation because it increases incidental exposure to novel information. Consequently, the use of social media leads to an increase in exposure to a wider range of political opinions than those normally encountered offline. The

findings of this research show evidence of polarisation and augmentation. The network analysis shows that there are tighter conceptual links within the anti-vaccination clusters by virtue of their more frequent and possibly highly selective interaction with scientific information extracted from open access journal articles (Van Schalkwyk, 2019). However, there is also evidence of article coupling spanning across groups who hold different views on the topic of vaccination. This could suggest exposure to novel information as suggested by Barberá (2015) and more complex patterns of interaction with scientific information by social movements (Dubois & Blank, 2018).

The findings show that some actors are more active than others while simultaneously spanning several communities by virtue of their topically broader collections of scientific articles. The concept of cognitive bridges as proposed in this paper provides a useful starting point for thinking about the possible roles that these individual actors play in communication networks, as well as the practical value of knowing who these bridging actors are.

Cognitive bridges were identified by mapping the interactions of actors with openly accessible sources of scientific information where the actors have self-organised into a social movement active on social media and the information source contains authoritative information that may be of value to the social movement and its cause. It may be tempting to describe the highly active and centrally located actors in the coupling network as intermediaries. But to do so would imply a certain degree of active interaction between actors and of intervention on the part of those actors (Frandsen, 2015). The only interaction in the tweeter coupling network is between actors and open access journal articles as actors extract information from those articles and insert the information into their tweets. In other words, the graphs and the results show communities of shared interests and highly active disseminators in those communities, but not necessarily actor interaction or their function as intermediaries.

Nevertheless, it is conceivable that while the tweeter coupling network may not represent actual tweeter interaction, the identification of certain types of actors may provide a useful starting point for isolating intermediaries in communication networks. This may present an opportunity to restore a polluted science communication environment or to disrupt nodes who contribute to its pollution. To illustrate, the identification of the cognitive bridge spanning opposing clusters in the network (Figure 5) offers the opportunity to verify the extent to which this tweeter intermediates in the online communications between anti-vaccination tweeters and pro-scientists, and to play a constructive role in online discussion and information sharing. The identification of a cognitive bridge between sub-clusters within the same social movement could point to intermediation of the kind that hardens beliefs across the movement and amplifies uncertainty in the broader communication network (Zannettou et al., 2017). Further research in such nodes may reveal and substantiate this possibility and may provide sufficient grounds for deleting such nodes from social media networks.

The findings of this paper may also call for a reconceptualization of the role of intermediaries in communication networks and how they may differ from traditional definitions of intermediaries. Intermediaries in communication networks may intermediate in ways that are less active or intentional. For instance, a simple retweet by a uniquely positioned tweeter can diffuse information much deeper into a network than a stream of tweets by a tweeter actively attempting to intervene in a specific information flow. Feng's (2016) notion of more passive "information bridges" versus the more active "influencers", "network builders" and "active engagers" may provide fertile ground for the development of the concept of intermediation in communication networks. As does research that shows that highly active provocateurs initiate conflict in online communication networks, but it is those less active who maintain the engagement between online communities in conflict (Kumar et al., 2018).

Limitations

Using the web as a site of study poses challenges. It is not uncommon for the owners of data sources on the web (e.g., Facebook and Twitter) to change the rules on how their data can be accessed and re-used. This may affect not only a researcher's ability to access data directly but may also compromise the functionality and usefulness of applications used to collect and analyse online data. There are also challenges with the stability of the data from online sources – websites disappear from the web for a variety of reasons, and users of social media close their accounts or change their account profiles (affecting either the content of those accounts or the accessibility of the account's content). These fluid conditions place limits on the replicability of this study and on the verifiability of selected data referenced in the research.

A further issue raised by Sugimoto et al. (2017) is the finding that discrepancies can exist when comparing results for mentions in the social media obtained from different aggregators. It is acknowledged that a limitation of this study is its dependency on data from Altmetric.com, and that the possibility exists that different results may be obtained when using any of the other available altmetric sources. Similarly, the findings could be made more robust by reducing the reliance on a single network graphing and analysis tool (i.e. NodeXL), and by introducing additional network analysis algorithms such as HITS or other random-walk based approaches to identify intermediaries and their relative positions in online communication networks.

Conclusion

The use of scientific information on a topic that has been settled by scientists but remains controversial in the public domain is not trivial. The World Health Organization has listed "vaccine hesitancy" (the reluctance or refusal to vaccinate despite the availability of vaccines) as one of 10 global health threats in 2019 (WHO, 2019). The anti-vaccination movement is strategic in its use of the social media to create and amplify uncertainty about vaccine safety in the broader public. Accessing information from scientific articles to bolster their claims forms part of the movement's communication strategy.

It has been shown that it is possible to create a tweeter coupling network using mentions on Twitter to open access scientific articles dealing with the publicly contested topic of vaccination. Using a coupling network based on altmetric data, this paper has shown how different tweeters can be located in the network, how distinct clusters are evident according to different ideological positions, and that tweeters link with each other to different degrees. This led to the observation of "super-disseminators" and the suggestion that there are tweeters who command "topical dominance" in the network. Ideologically separated "communities of shared interests" are shown to exist on Twitter and certain actors are more active than others while simultaneously spanning several communities by virtue of their topically broader collections of scientific articles. These tweeters are described as "cognitive bridges" and may play important roles in the distribution of information in communication networks. Managing or controlling information flows may depend on these important actors.

This paper has illustrated how tweeter coupling and network analysis can be applied to developing a better understanding of the communication of science in social media. Additional empirical research on pathways, intermediation, and the multiplicity of roles and characteristics of actors within online social movements is needed to further contribute to a more fine-grained understanding of the potentials of open science and, more importantly, how science can respond to the effects of greater openness and socially networked communication.

Acknowledgements

This research is partially funded by the South African DST-NRF Centre of Excellence in Scientometrics and Science, Technology and Innovation Policy (SciSTIP) and the South African DST/NRF Research Chair in Science Communication, Stellenbosch University.

References

Archambault, E., Amyot, D., Deschamps, P., Nicol, A., Provencher, F., Rebout, L., & Roberge, G. (2014). Proportion of Open Access Papers Published in Peer-Review Journals at the European and World Levels 1996-2013. Brussels: European Commission. Retrieved from http://science-metrix.com/sites/default/files/science-metrix/publications/d_1.8_sm_ec_dg-rtd_proportion_oa_1996-2013_v11p.pdf

Barberá, P. (2015). How social media reduces mass political polarization: Evidence from Germany, Spain, and the US. Paper presented at the 2015 APSA Conference, 3-6 September, San Francisco. Retrieved from http://tinyurl.com/kyrqfe7

Bar-Ilan, J., Haustein, S., Peters, I., Priem, J., Shema, H., & Terliesner, J. (2012). Beyond citations: Scholars' visibility on the social web. *17th International Conference on Science and Technology Indicators* (pp. 98-109). Montreal: Science-Metrix.

Bean, S. (2011). Emerging and continuing trends in vaccine opposition website content. *Vaccine, 29*, 1874-1880.

Bennato, D. (2017). The shift from public science communication to public relations: The Vaxxed case. *Journal of Science Communication, 16*(2), 1-11.

Blöbaum, B. (2016). Key factors in the process of trust: On the analysis of trust under digital conditions. In B. Blöbaum, *Trust and Communication in a Digitized World: Models and concepts of trust research* (pp. 3-26). Heidelberg: Springer.

Bucchi, M. (2018). Credibility, expertise and challenges of science communication 2.0. *Public Understanding of Science, 26*(8), 890-893. doi:10.1177/0963662517733368

Burt, R. (2001). Structural holes versus network closure as social capital. In N. Lin, K. Cook, & R. Burt, *Social Capital: Theory and research* (pp. 31-56). New York: Aldine De Gruyter.

Castells, M. (2009). *Communication Power.* Oxford: Blackwell.

Castells, M. (2017). Rethinking development in the global information age. In J. Muller, N. Cloete, & F. van Schalkwyk, *Castells in Africa* (pp. 67-94). Cape Town: African Minds.

Costas, R., De Rijcke, S., & Marres, N. (2017). Beyond the dependencies of altmetrics: Conceptualizing 'heterogeneous couplings' between social media and science. Paper presented at *Altmetrics17: The dependencies of altmetrics*. Toronto. Retrieved from Altmetrics: http://altmetrics.org/wp-content/uploads/2017/09/altmetrics17_paper_4-1.pdf

Dickel, S., & Franzen, M. (2016). The 'problem of extension' revisited: New modes of digital participation in science. *Journal of Science Communication*, 15(1), 1-15. doi:10.22323/2.15010206.

DiResta, R., & Lotan, G. (2015, June 8). Anti-vaxxers are using Twitter to manipulate a vaccine bill. *Wired.* Retrieved from https://www.wired.com/2015/06/antivaxxers-influencing-legislation/

Dubois, G. & Blank, G. (2018). The echo chamber is overstated: the moderating effect of political interest and diverse media. Information, Communication & Sociey, 21(5), 729-745. doi: 10.1080/1369118X.2018.1428656.

Feng, Y. (2016). Are you connected? Evaluating information cascades in online discussion about the #RaceTogether campaign. *Computers in Human Behavior*, 54, 43-53. doi: 10.1016/j.chb.2015.07.052.

Frandsen, F. & Johansen, W. (2015). Organizations, stakeholders, and intermediaries: Towards a general theory. *International Journal of Strategic Communication*, 9(4), 253-271. doi: 10.1080/1553118X.2015.1064125.

Hansen, D., Shneiderman, B., & Smith, M. (2010). *Analyzing Social Media Networks with NodeXL: Insights from a connected world.* Burlington MA: Morgan Kaufman.

Haustein, S., & Costas, R. (2015). Identifying Twitter audiences: Who is tweeting about scientific papers? *ASIS&T SIG/MET Metrics 2015 Workshop* (pp. 1-3).

Haustein, S., Costas, R., & Lariviere, V. (2015). Characterising social media metrics of scholarly papers: The effect of document properties and collaboration patterns. *PLOS ONE, 10*(5), e0127830. doi:10.1371/journal.pone.0120495.

Himelboim, I., Smith, M., Raine, L., Shneiderman, B., & Espina, C. (2017). Classifying Twitter topic-networks using social network analysis. *Social Media + Society*, 1-13. doi:10.1177/20563051176915.

Jasanoff, S. (2006). Transparency in public science: Purposes, reasons, limits. *Law and Contemporary Problems, 69*(21), 21-45.

Kahle, S. (2011). Preferential attachment. In A. Barnett, *Encyclopedia of Social Networks* (pp. 718-720). Thousand Oaks: Sage.

Kata, A. (2012). Anti-vaccine activists, Web 2.0, and the postmodern paradigm: An overview of the tactics and tropes used online by the anti-vaccination movement. *Vaccine, 30*, 3778-3789.

Kessler, M.M. (1963). Bibliographic coupling between scientific papers. *American Documentation*, 14, 10-25. doi:10.1002/asi.5090140103

Kumar, S., Hamilton, W., Leskovec, J., & Jurafsky, D. (2018). Community interaction and conflict on the web. WWW 2018: The 2018 Web Conference, April 23-27, 2018, Lyon, France (pp. 1-11). New York: ACM.

Landrum, A. (2017). The role of intermediaries in communicating science: A synthesis. In K. Jamieson, D. Kahan, & D. Scheufele, *The Oxford Handbook of the Science of Science Communication* (pp. 253-260). New York: Oxford University Press.

Leask, J. (2015). Should we do battle with anti-vaccination activists? *Public Health Research & Practice, 25*(2), e2521515. doi:10.17061/phrp2521515.

Leonelli, S., Spichtinger, D., & Prainsack, B. (2015). Sticks and carrots: Encouraging open science at its source. *Geography and Environment, 2*, 12-16. doi:10.1002/geo2.2.

Miller, V. (2017). Phatic culture and the status quo: Reconsidering the purpose of social media activism. *Convergence*, 23(3), 251-269. doi:10.1177/1354856515592512.

Muller, J. (2017). Universities and the 'new society'. In J. Muller, N. Cloete, & F. Van Schalkwyk, *Castells in Africa: Universities and Development* (pp. 17-31). Cape Town: African Minds.

Ortiz-Ospina, E., & Roser, M. (2016). Trust. *Our World in Data*. Retrieved from: https://ourworldindata.org/trust

Piwowar, H., Priem, J., Larivière, V., Alperin, J., Matthias, L., Norlander, B., . . . Haustein, S. (2017). The state of OA: A large-scale analysis of the prevalence and impact of open access articles. *PeerJ Preprints* 5, e3119v1. doi:10.7287/peerj.preprints.3119v1.

Schäfer, M. (2017). How changing media structures are affecting science news coverage. In K. Jamieson, D. Kahan, & D. Scheufele, *The Oxford Handbook of the Science of Science Communication* (pp. 51-59). New York: Oxford University Press.

Scheufele, D. (2013). Communicating science in social settings. *PNAS, 10*(3), 14040-16047.

Scheufele, D. (2014). Science communication as political communication. *PNAS, 111*(4), 13585-13592.

Southwell, B. (2017). Promoting popular understanding of science and health through social networks. In K. Jamieson, D. Kahan, & D. Scheufele, *The Oxford Handbook of the Science of Science Communication* (pp. 223-231). New York: Oxford University Press.

Stalder, F. (2006). *Manuel Castells: The theory of the network society.* Cambridge: Polity.

Sugimoto, C., Work, S., Larivière, V., & Haustein, S. (2017). Scholarly use of social media and altmetrics: A review of the literature. *Journal of the Association for Information Science and Technology*, 68(9), 2037-2062. doi:10.1002/asi.23833.

Thelwall, M., Tsou, A., Weingart, S., Holmberg, K., & Haustein, S. (2013). Tweeting links to academic articles. *Cybermetrics: International Journal of Scientometrics, Informetrics and Bibliometrics,* 17, 1-8.

Tucker, J., Guess, A., Barberá, P., Vaccari, C., Siegel, A., Sanovich, S., . . . Nyham, B. (2018). *Social Media, Political Polarization, and Political Disinformation: A review of the scientific literature*. Washington DC: William and Flora Hewlett Foundation.

Van Noorden, R. (2015). Online collaboration: Scientists and the social network. *Nature, 512*(7513), 126-129. doi:10.1038/512126a.

Van Schalkwyk, F. (2019). New potentials in the communication of open science with non-scientific publics: The case of the anti-vaccination movement Doctoral dissertation. Stellenbosch University. doi: 10.13140/RG.2.2.28501.06888

Weingart, P. (2011). Science, the public and the media – Views from everywhere. In M. Carrier, & A. Nordmann, *Science in the Context of Application* (pp. 337-348). Dordrecht: Springer.

WHO (World Health Organization) (2019). Ten threats to global health in 2019. WHO website. https://www.who.int/emergencies/ten-threats-to-global-health-in-2019

Williams, J. (2018). *Stand Out of My Light: Freedom and resistance in the attention economy*. Cambridge: Cambridge University Press.

Yeo, S., Xenos, M., Brossard, D., & Scheufele, D. (2015). Selecting our own science: How communication contexts and individual traits shape information seeking. *The ANNALS of the American Academy of Political and Social Science*, 658(1), 172-191. doi:10.1177/0002716214557782.

Zannettou, S., Caulfield, T., De Cristofaro, E., Kourtellis, N., Leontiadis, I., Sirivianos, M., ... Blackburn, J. (2017). *The web centipede: Understanding how web communities influence each other through the lens of mainstream and alternative news sources*. IMC17, November 1–3, 2017, London, UK. New York: ACM. doi:10.1145/3131365.3131390.

Zhao, D., & Strotmann, A. (2008). Evolution of research activities and intellectual influences in information science 1996 – 2005: Introducing author bibliographic-coupling analysis. *Journal of American Society for Information Science and Technology*, 59, 2070-2086. doi:10.1002/asi.20910

Zimmerman, R., Wolfe, R., & Fox, D. (2005). Vaccine criticism on the world wide web. *Journal of Medical Internet Research, 7*(2).

Gender disparities in the field of economics

Liu Junwan [1], Song Yinglu [1], Yang Sai [1], Cassidy r. Sugimoto [2], Vincent Lariviere [3]

[1] *liujunwan@bjut.edu.cn 18810621725@163.com 850846582@qq.com*

School of Economics and Management, Beijing University of technology, Beijing, China

[2]*sugimoto@indiana.edu*

School of Informatics and Computing, Indiana University Bloomington，USA

[3]*vincent.lariviere@umontreal.ca*

École de bibliothéconomie et des sciences de l'information, Université de Montréal, Canada

Abstract

In order to analyse the gender disparities in scientific research output in the field of economics, this paper selected the Web of Science database as the source database. We collected and screened 257,642 articles written by 130,397 authors from 1933 to 2017 in the field of economics. In this study, we use mathematical statistics and bibliometrics indexes to quantitatively analyse the gap between male and female authors in many aspects, including the output and influence in different level of journals and institutions, the dynamic evolution of output and influence and cooperation modes with gender disparities. In addition, we have analysed the disparities in output and influence of male and female authors among different countries. The results show that male authors dominate in the economics research field according to their high output and influence. However, female authors also show advantage when it comes to the research influence. This study can provide an insight of gender different in economics research.

Keywords: gender disparities, economics, output

Introduction

Gender disparities are not only reflected in biological structure and social roles. In the academic field, gender disparities also exist widely and have been influencing the quantity and quality of academic activities for a long time. With the development of the feminist movement and the progress of society, the barriers preventing women from entering the scientific community are gradually being eliminated, but the negative effects of gender role positioning and education restriction still exist, and gender is still an important factor affecting the balanced development of scientific research.

In recent years, many scholars have studied gender disparities in science. Shi Yuantao and Chen Xueling (2011) investigated and analysed the scientific and technological personnel in institutions of higher learning and scientific research institutes in Hubei province and found that the proportion of men in the groups with high scientific research achievements was much higher than the proportion of women. Lin (1997) surved 441 Chinese scientists and also found that women make up a larger proportion of authors with lower scientific production than men. Ma Ying, Zhao Yanling and Gong Xin (2018) pointed out that only 6% of academics were female, and 87% of female college students faced gender discrimination during their job search. Yuan Yuzhi (2017) found that family burden and the level of participation in cooperative scientific research were the key variables influencing the output of scientists. Zhang Jinjie and Zhang Dongshuo (2005) described the phenomenon of the "relative absence" of women in the natural sciences, which was mainly reflected in the obvious gender gap

between female and male scientists in the number, level, fields and achievements of scientific research. The traditional concept of the social division of labour and women's personal reasons contribute to this phenomenon.

Schiebinger and Linda (2014)suggested that gender must be taken into account in scientific research, especially in the biomedical field. Guglielmi and Giorgia (2018) demonstrated that men were more successful than women in applying for research grants. The study by Yu Xie and Shauman (1998) found that although the difference between the number of female scientists and male scientists was gradually narrowing, men still ranked higher than women in the scientific community. Preston A (1994) pointed out that women's investment and participation in scientific research are increasing, but women are twice as likely as men to leave science after graduating from school. Garg (2014) calculated the contribution of female authors to papers, showing that female scientists were slightly less productive than men at the individual level. Sax and Hagedorn (2002) showed that in the 30 years from 1972 to 1999, both male and female teachers' scientific research output increased, but the gender gap among high-yielding teachers remained unchanged. Cassidy R. Sugimoto and Vincent Lariviere(2013) presented a bibliometric analysis that confirmed that gender imbalance still exists in research output worldwide. Kyvik and Teigen (1996) investigated the scientific research output of teachers in four Norwegian universities and found that there was also a gender difference in the relationship between age and productivity, which was that the gender difference decreased at first and then increased as age increased. Rauber and Ursprung (2007) confirmed this conclusion, and they found that the scientific research output of female economists changed regularly with the development of their careers. Sotudeh H (2014) showed that although there were only a few female researchers in the field of nanotechnology, the women researchers were also more likely to publish in high-impact journals and performed well in terms of scientific achievements and influence. Mauleo’n and Bordons’s(2006) research results showed that there were no significant disparities in scientific output and influence compared with men and women in the field of materials science and that women tend to publish in high impact journals.

A review of these literature shows that gender difference is an important factor that affects the output and influence of female scientists in scientific research in both developed and developing countries. Diversity and equality are needed in the field of scientific research, so gender-specific assessments of scientific achievements will demonstrate different scientific capacities and thus contribute to the development of strategic plans that enable women to develop themselves in the scientific community.

This paper takes the field of global economic research data as an example. The main purpose of the paper is to quantitatively analyse the output of scientific research and influence disparities between male and female scientists, preliminarily explore the influence of gender disparities on scientific contributions and comprehensively investigate the gender disparities in scientific research activities in terms of the length of author careers, institutions, nationalities and other aspects. The research results can provide quantitative data and reference for science and technology decision makers to make policy, and make academia develop more balanced.

Data and methods

To ensure the representativeness and authority of data, this paper selected the Web of Science database as the source database. We collected and screened 257,642 articles written by 130,397 authors from 1933 to 2017 in the field of economics. The field of economics was defined using the NSF fields and subfields classification of journals. A total of 450,566 papers published in 348 economics journals were analysed in this study. Authors were disambiguated using the methods developed by Caron and van Eck (2014). The gender of authors was assigned using authors' first names, following the method developed by Larivière et al. (2013). This paper mainly adopts the mathematical statistics method that is commonly used in bibliometrics to conduct regression analysis and the non-parametric test for the collected data with the aim to investigate the quantitative relationship between scientific research productivity and career. At the same time, an independent sample T test was conducted on gender disparities in scientific research output and on the influence of male and female authors to verify whether there were significant difference in the distribution of variables between the two groups. The tools used were mainly Excel for basic data statistics, SPSS for regression analysis and a non-parametric test and Gephi for visualization analysis to draw the network diagram of scientific research cooperation in the field of economics, and node centrality was calculated.

This paper adopts the average counting method to calculate the number of papers (regardless of the number of collaborators). The average counting method (Pang Jingan, 1999) refers to the number of papers calculated according to the method of "one per person" regardless of the rank of authors.

Average annual publication (AAP) and career length (CL): average annual publication refers to the number of publications per year, which is equal to the total output/career length. Career length refers to the time interval between the publication of the author's first paper and the publication of his or her last paper to date.

$$\text{AAP} = \text{Total number of annual publications/CL} \tag{2.1}$$

Average annual citation number (AAC) and paper citation frequency (PCF): average annual citation number refers to the average annual citation number of a single paper, and paper citation frequency refers to the average annual citation number of all articles of an author, which are added together and averaged to measure the influence of an author. The calculation formula is as follows:

$$\text{AAC} = \text{Total number of annual citations in Y year/(2017-Y+1)} \tag{2.2}$$

$$\text{PCF} = \text{AAC/Total number of paper} \tag{2.3}$$

Paper contribution rate: as a paper is often completed by multiple authors, the total of n papers are published by an author, and the number of authors for each paper is mi. Therefore, the author's paper contribution rate formula is:

$$\sum_{i=1}^{n} \frac{1}{m_i} / n \tag{2.4}$$

Cooperative network graph: to show the cooperative relationship between authors, each author can be abstracted into a node in the graph, and a line can be connected between the two authors (two nodes) to generate the cooperative network graph.

Degree centrality: the degree centrality of a node is also referred to as the degree of the node, which is defined as the total number of edges that the node connects with other nodes.

The calculation formula is as follows:

$$C_D(N_i) = \sum_{j=1}^{g} x_{ij}\ (i \neq j)$$

Concentration index: the concentration index is a measure of the correlation between two variables based on frequency data that varies around the neutral value 1. The index refers to the ratio of the proportion of a certain class of data in the new category to the proportion of the original class in the total sample after classification according to the new classification method. In this paper, the concentration index refers to the ratio of a gender in different groups to the ratio of that gender among all authors.

Gender disparities in output and influence

Of the 130,397 authors, the proportion of men is 2.45 times higher than that of women, indicating that the number of male scholars in the field of economics is far higher than that of female scholars. The first authors of all the papers were counted, among which 81% were men and 19% were women. Statistical data are shown in Table 1: Part of the reason for the large gap between men and women is that the unequal status of men and women causes men to have more opportunities to participate in academic discussions and gain more recognition, while women are less likely to participate in academic discussions because they are more likely to be questioned. Without those opportunities, women can miss out on research collaborations and job offers.

Table 1 Ratio of male to female authors

		M	*F*	*TOTAL*
All the authors	Author Num(proportion)	92633 (71%)	37764 (29%)	130397
The first author	Paper Num (proportion)	186512 (81%)	43159 (19%)	229671

Research on gender disparities in output and influence

Considering that the number of male authors is far higher than the number of female authors, so AAP is adopted instead of the total number to measure the output of each author. By calculating the overall data, male authors' AAP was 0.67, and female authors' AAP was 0.66, making the latter slightly lower than the former. An independent sample T test was used to verify whether the gender difference in output is significant.

Table 2 Independent sample T test of scientific research output

The set of statistics					
	Gender	*N*	*The mean*	*The standard deviation*	*Standard error of the mean*
AAP	F	37730	.663759	.4415452	.0022732
	M	92622	.669698	.4803193	.0015782

Independent sample test								
		Levene test of variance equation		*T test for the mean equation*				
							95% confidence intervals	
		F	*Sig.*	*t*	*Sig. (double side)*	*The mean difference*	*Lower limit*	*Upper limit*
AAP	Variance is equal	112.610	.000.	-2.071	.038.	-.0059385	-.0115578	-.0003193
	Variance is not equal			-2.146	.032.	-.0059385	-.0113625	-.0005146

According to the Levene test results of the variance equation in Table 2, the F value is 112.61, and the sig value is 0.000<0.05, indicating that the difference of the variance between the two groups is significant. In the T test in Table 2, sig (both sides) =0.032<0.05, that is, there is a significant difference of the mean in the two groups, indicating that the difference in the scientific research output between male and female authors is significant.

PCF is used to measure the influence of scientific researchers. By calculating the overall data, it was concluded that the male authors' PCF is 1.48 and the female authors' PCF is 1.28, the latter still being slightly lower than the former. Again, an independent sample T test was used to verify whether the gender difference in influence is significant.

Table 3 Independent sample T test of scientific research influence

The set of statistics					
	Gender	*N*	*The mean*	*The standard deviation*	*Standard error of the mean*
PCF	M	360374	1.481720	3.6044713	.0060043
	F	90089	1.284815	2.7894445	.0092936

Independent sample test								
		Levene test of variance equation		*T test for the mean equation*				
							95% confidence interval	
		F	*Sig.*	*t*	*Sig. (double side)*	*The mean difference*	*Lower limit*	*Upper limit*
PCF	Variance is equal	422.513	.000	15.292	.000	.1969052	.1716674	.2221429
	Variance is not equal			17.796	.000	.1969052	.1752191	.2185912

According to the Levene test results of the variance equation in Table 3, the F value is 422.513, and the sig value is 0.000<0.05, indicating that the difference of the variance between the two groups is significant. Sig (both sides) =0.000<0.05 in the T test in Table 3, that is, there is a significant difference between the means of the two groups, indicating that the difference in the scientific research influence between male and female authors is significant.

Gender disparities among high-product and high-impact authors

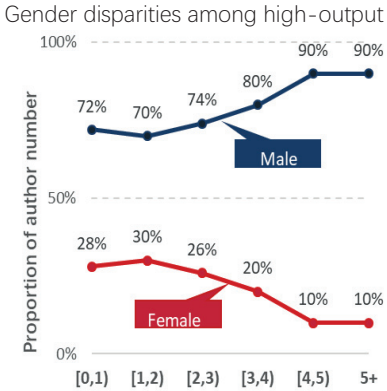


Figure 1 Proportion of author number in different annual output ranges

The average annual publication of authors is divided into 6 intervals from small to large, and the proportion of female authors in each interval is shown in Figure 1. It can be seen that with an increase in average annual publications, the proportion of female authors continues to decrease. The higher the productivity, the greater the gender difference between authors.

In considering the number of authors and the average annual output of all authors, this paper defines the authors with an average annual publication of more than 2, (1-2) and (0-1) as high-product, medium-product and low-product

authors, respectively, which account for 2%, 44% and 54% of the total population, respectively. To explore and compare the gender disparities within the three groups, the concentration index was introduced to represent the correlation between the proportion of male and female authors in a certain group and the proportion of male and female authors in all the groups.

Table 4 Proportion of male to female authors among the three groups

		F	*M*
High-product 2450	Number	514 (21%)	1936 (79%)
	Concentration index	0.72	1.11
Medium-product 57417	Number	17196 (30%)	40221 (70%)
	Concentration index	1.03	0.99
Low-product 70530	Number	19947 (28%)	50583 (72%)
	Concentration index	0.97	1.01

As seen in Table 4, the concentration index of male and female authors with low and medium production changes around the median value of 1, and there is no significant difference. The concentration index of males among high-product authors was 1.11, which was approximately 1.5 times higher than that of female high-product authors. In other words, the number of male authors was more concentrated among the high-product authors.

Gender disparities among high-impact authors

As mentioned before, paper citation frequency is used to measure the influence of scientific researchers by calculating the average citation frequency of all papers published by each author and dividing the citation frequency into 5 intervals from small to large. As shown in Figure 2, there is no significant change in the proportion of female authors in different cited frequency intervals, which is approximately 29%.

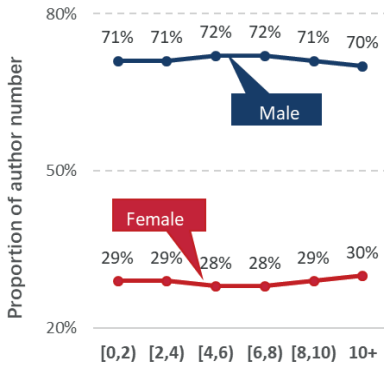


Figure 2 Proportion of author numbers in paper citation frequency ranges

By considering the number of citations of an article and the author citation frequency, this paper defines the authors with a paper citation frequency of more than 6, (1-6] and (0-1] as high-impact, medium-impact and low-impact, respectively, which account for 2%, 30% and 68% of the total population, respectively. To explore the internal gender disparities within each of the three groups and to compare the gender disparities between the three groups, we continue to use the concentration index

Table 5 Proportion of male to female authors among the three groups

		F	*M*
High-impact 2589	Number	739 (29%)	1850 (71%)
	Concentration index	1.00	1.00
Medium-impact 39693	Number	11487 (29%)	28206 (71%)
	Concentration index	1.00	1.00
Low-impact 88115	Number	25538 (29%)	62577 (71%)
	Concentration index	1.00	1.00

It can be seen from Table 5 that there is no significant difference in the concentration of

men and women in the three groups, as the concentrations of both change around the median value of 1. However, we are surprised to find that in the high-impact group, the articles of female authors are cited 10.15 times, which is higher than the citation rate of male authors in that group (9.66 times), while the number of female authors in the high-impact author group is only two-fifths of the number of male authors, which reflects the female authors' abilities and advantages in scientific competition.

Gender disparities at different levels of journals and institutions

Gender disparities at different levels of journals

Journal impact factor is an important index for the quantitative evaluation of journals that was first put forward by the founder of SCI in the United States (E. Garifedl) and has been widely used to evaluate the academic level of journals (Jin Bihui, Wang Shouyang, 1999). The greater the impact factor of the journal, the greater the influence of the journal is. In this paper, journals are arranged according to the order of influencing factors from high to low. The top 10%, 10%-25%, 25%-50% and 50% and above journals are defined as first-level, second-level, third-level, and fourth-level journal, respectively. As shown in Figure 3, the number of papers published by male authors is higher than that of female authors in journals across all levels.

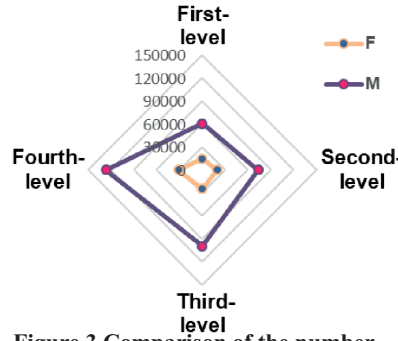


Figure 3 Comparison of the number of papers published in different levels

Gender disparities in different levels of institutions

We rank the total citation frequency of articles published by each academic institution from large to small and define the top 10% most frequently cited institutions as core institutions and the rest as non-core institutions in this field. As shown in Figure 4, the number of female authors in the core institutions accounts for approximately 26% and that of male authors accounts for approximately 74%. The number of female authors in non-core institutions accounts for approximately 30% and that of male authors accounts for approximately 70%. Compared with the proportion of male authors, it can be seen that there are fewer female authors in the core institutions.

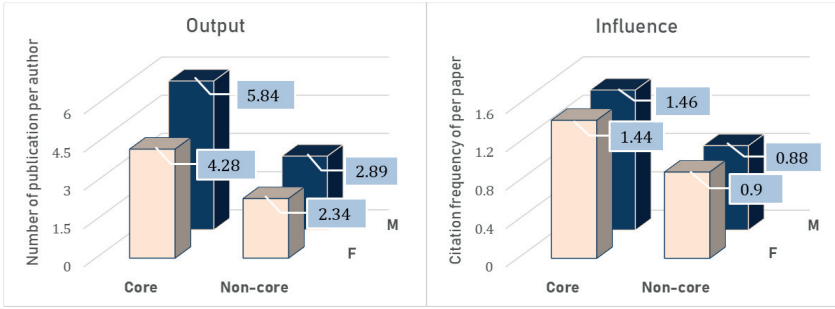


Figure 4 Gender disparities in output and influence

It can be seen from Figure 4 that the output of female authors is also significantly different from that of male authors at institutions of the same level, and the difference at core

institutions is greater. In terms of influence, the influence of male authors is slightly higher than that of female authors at core institutions, while that of female authors is slightly higher than that of male authors at non-core institutions.

Dynamic evolution of output and influence and gender difference in Top10 countries

To further explore the dynamic evolution law of output and influence of male and female authors with career development, we first counted the average career length of male authors and female authors. The average career length of males is 2.64 years longer than that of females. Male authors can have careers of up to 73 years compared with 39 years for female authors. Almost all of those whose career length is more than 40 years are Americans.

Dynamic evolution of output and influence

The output of male and female authors at different stages of their careers was studied by comparing the number of publications and using SPSS for the regression analysis of the dynamic evolution of output, as shown in Figure 5(a). The number of publications by women in SPSS fitting is y1 and the female career is x1 in the curve equation of $y_1 = 9376.25 + 2520.52 * \ln(x_1)$. The curve equation between the number of male publications (y_2) and male career (x_2) is $y_2=27850.01 + -7105.97 * \ln(x_2)$. Both fitting curves are logarithmic functions with negative logarithmic coefficients, which indicates that the output of male and female authors decreases logarithmically with the growth of career age.

The influence of male and female authors at different stages of their careers was studied by comparing the citation frequency of each paper. Due to the lack of papers published by female authors after 38 years of a career and the few papers published by male authors, the fluctuation of citation frequency is so large that the results are not representative. Therefore, only the influences of papers published in the first 38 years of a career are compared here.

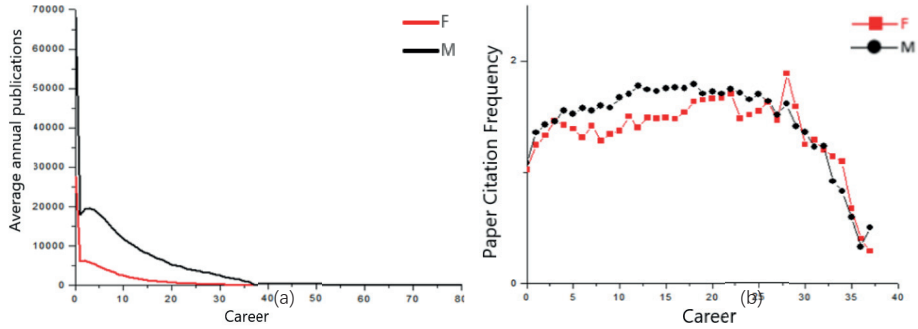


Figure 5 The dynamic evolution of the output(a) and the influence(b)

As shown in Figure 5 (b), the influence of male and female authors shows a trend of fluctuation with the development of the career; on the whole, men and women reach peak influence approximately 20 years into their careers, and their influence declines rapidly in the next fifteen years. However, the influence of female authors is significantly higher than that of male authors in the first 28 years of a career, and after that point, the influence of female authors is slightly higher than that of male authors, which is very unexpected. Part of the reason for this phenomenon is that women have higher qualifications and gradually gain more influence in the later stages of their careers. It indicates that women still have a higher

influence even though the number of women is relatively small.

The proportion of female authors in the TOP10 countries

This section expands on the analysis of section 4.1 to further explore the gender disparities of output and influence in different countries. The number of papers published in all countries was calculated in this study. To make the data more representative, only the papers published in the top 10 countries were statistically studied for gender disparities. These ten countries are the United States, Great Britain, Germany, Canada, Spain, Australia, Italy, Italy, France, the Netherlands and China, and they are referred to as the TOP10 countries henceforth.

We firstly count the proportion of men and women in the TOP10 countries. The result shows that there are more male authors than female authors in all countries, and the proportion of female authors is between 20% and 40%. Spain and Italy have the highest proportion of female authors at 38%, and China has the lowest proportion of female authors at 24%.

Gender disparities in output and influence of the TOP10 countries

As shown in Figure 6, the output of men in the TOP10 countries is more than twice that of women. The output of American male and female authors is significantly higher than that of authors from other countries. Outside the USA, there is a small gap between the influence of male and female authors in the rest of the nine countries.

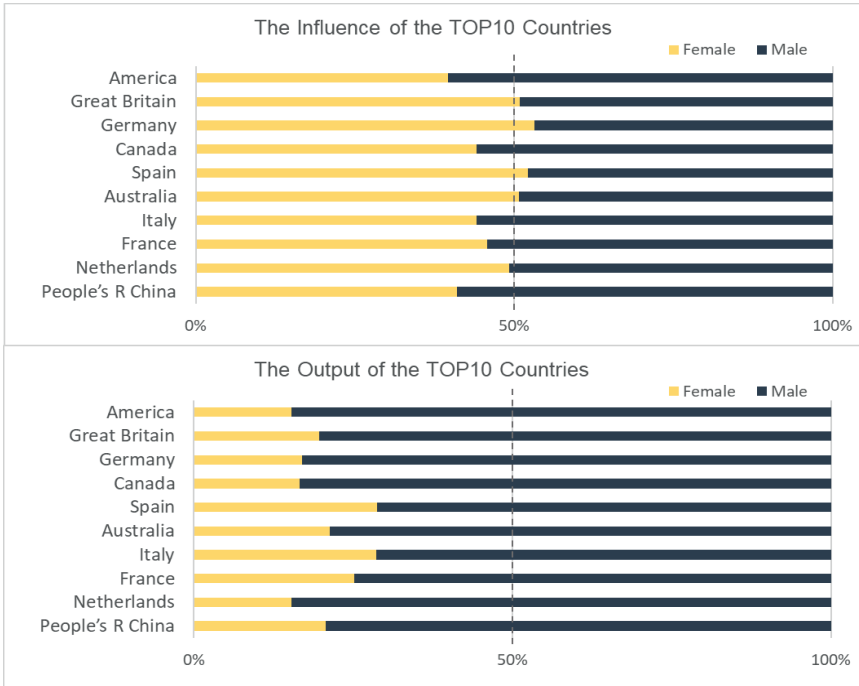


Figure 6 Output and influence of the TOP10 countries

In Great Britain, Germany, Spain, and Australia, the influence of male authors is even lower than that of female authors, although the number of published male authors is almost twice as much as the number of published female authors. Although the number of women is

only about a quarter of the number of men, the influence of female authors' achievements is relatively outstanding in the field of scientific research. This result is consistent with the result in different institutions.

Research on the cooperation mode with gender disparities

Analysis of centrality in a cooperative network

This paper uses degree centrality as an indicator to measure the importance of authors in cooperative networks (Liu Zhi peng, Zeng Yi, Wang Ting, 2014). Scholars with a high degree of centrality are considered important central people in the cooperative network. The average centrality of female authors was 2.90 and that of male authors was 3.40. This indicates that male authors have a higher influence than female authors in the network of scientific cooperation in the field of economics. The top 10 authors in the centrality ranking are all male, and their centrality is higher than 100. The highest author has a centrality of 204, which is enough to prove that male authors with leading positions in the cooperation network play an important role in discipline construction and information dissemination in the field of economics.

In addition, author's paper contribution rate is another way to measure scientific research output. The paper contribution rates of male and female authors were calculated according to formula 2.3. The average paper contribution rate of female authors is 0.47, which means that the output per female author is approximately 0.47 articles. The average paper contribution rate of male authors is 0.51, which means that the output per male author is approximately 0.51 articles. It is concluded that the average contribution of male authors is higher than that of female authors.

Analysis of output and influence in different cooperation modes

The output and influence of different gender combinations were analysed. If dividing the combination into three forms, MM indicates male cooperation—that is, if all authors of a paper are male; FF indicates female cooperation—that is, all authors of a paper are female; FM indicates at least one male and one female author of a paper. The proportion of each combination is shown in Figure 7. Thus, male authors' communication dominates the cooperation network of economic research. Men have obvious advantages in establishing relations with those in the core academic field, which has long been dominated by men.

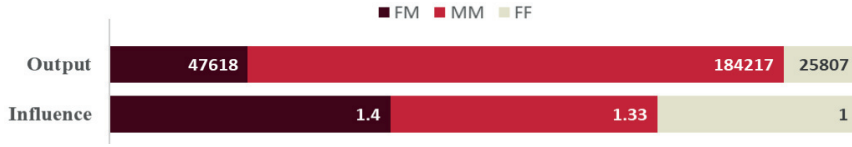


Figure 7 Disparities in output and influence under different group combinations

Comparing the average citation frequency of the articles, it can be seen that the paper citation frequency in the FM group can reach 1.40, which means the papers published by FM groups are more likely to be cited and have the highest influence. The MM group, with a paper citation frequency of 1.33, follows the FM group closely, and the paper citation frequency in the FF group is 1.00. This suggests that the participation of the opposite sex in a team is conducive to improving the influence of scientific research results. Therefore, it is

necessary to encourage more female authors to participate in scientific cooperation and establish cooperative relations between male and female researchers.

Conclusion

This paper makes a powerful quantitative analysis of the gap between male and female authors from such aspects as the number of male and female authors, output and influence of different journals and institutions, career, dynamic evolution of output and influence, and cooperation mode of gender difference.The results show that male authors dominate in the scientific research field because of their high output and influence. However, although the number of female authors is only about a quarter of the number of male authors, female authors do not lag behind and even score higher than male authors in regard to the influence of their scientific research. In addition, the influence of research results is enhanced by the participation of women. As a result, the important role of female scientists in scientific research cannot be ignored. The significant gap between male and female authors in science is reflected in every country, but the gap is relatively smaller in places such as Europe and South America, which is due to a series of policies that encourage women to participate in scientific and technological activities. Academia needs to focus on diversity in academia and give scholars more equal opportunities. Countries should work together to adopt policies that support families so as to reduce women's family burdens and provide more opportunities for women to realize self-worth, which will be of great help in arrowing the status gap between men and women and in the balanced development of entire scientific research fields. This paper only studies the gender difference in the field of economics and has not covered all disciplines, so the research results have some limitations. Since the distribution of scholars is closely related to academic characteristics, in order to deeply explore the disparities between male and female scholars, data from other disciplines should also be analyzed in future work, which will provide more comprehensive data support for science and technology policymakers.

Acknowledgments

This research received the financial support from National Science Foundation of China: Study on the Structure and Evolution of the Scientific Collaboration Network of Academicians from the Perspective of Symbiosis: A Case Study of Academicians of CAS and NAS under grant number 71603015. Also supported by the Natural Science Foundation of Beijing, China (Grant No. 9182001). Our gratitude also goes to the anonymous reviewers for their valuable comments.

References

Caron, E., & van Eck, N. J. (2014). Large scale author name disambiguation using rule-based scoring and clustering. In Proceedings of the 19th international conference on science and technology indicators (pp. 79-86).

Danell, r. & Hjerm, m. (2012). Career prospects for female university researchers have not improved. *Scientometrics*. 94, 999-1006.

Fox, m. f. Gender. (2005). Family characteristics and publication productivity among scientists. *Social Studies of Science*. 1(35), 131-150.

Garg K C,Kumar s. (2014). Boston profile of Indian scientific output in life sciences with a focus on

the contributions of women scientists. *Scientometrics*. 98, (pp.1771-1783).

Guglielmi, Giorgia. (2018). Gender bias tilts success of grant applications. *Nature*. 554(7690), (pp.14-15).

Jin bihui, Wang Shouyang, Ren Shengli, Liu Yajuan. (1999). On the relationship between journal influencing factors and academic quality of papers. *Monograph and review*. 11(4), (pp.202-205).

Larivière, V., Ni, C., Gingras, Y., Cronin, B., & Sugimoto, C. R. (2013). Bibliometrics: Global gender disparities in science. Nature, 504(7479), (pp.211-213).

Li Lexuan, Wen ke. (2008). Review and enlightenment of foreign policies and measures to promote women's participation in science and technology. *Journal of China women's university*. 20(6), (pp.75-80).

Lin Ju-ren. (1997). Introduction to research on gender differentiation in science. *Collection of women's studies*. (2), (pp.49-53).

Liu Zhipeng, Zeng yi, Wang ting. (2014). Coauthor network data model construction and core author mining research. *Software guide.* 12, (pp.141-143).

Ma Ying, Zhao Yandong, Gong Xu, Sun Li, Zheng Yonghe. (2018). Close the gender gap in Chinese science. *Nature*. 557(7703), (pp. 25-27).

Mauleo ´ n, E. & Bordons, m. (2006). The Productivity, impact and publication habits by gender in the area of materials science. *Scientometrics.* 66(1), (pp.199-218).

Pang Jingan. (1999). Research methods of scientific metrology. *Beijing: science and technology literature publishing house.*

Preston A E (1994). Why have all the women gone? A study of exit of women from science and engineering professions. *American Economic Review.* 84, (pp.1446-1462).

Rauber, M. & Ursprung, H.W. (2007). Life Cycle and Cohort Productivity in Economic Research: The Case of Germany. *CESifo Working Paper.* 9(4), (pp.2-46).

Sax, L. J. & Hagedorn, l. s.,& Arredondo, m.,& Dicrisi F.A.R. (2002). Faculty Research Productivity Exploring the Role of Gender and family-related factors. *Research in Higher Education.* 43(4), (pp.423-446).

Schiebinger, Londa. (2014). Scientific research must take gender into account. *Nature.* 507(7490), (pp.9-9).

Shi Yuantao, Chen Xueling. (2011). Exploration of the mystery of scientific research output of female scientists from the perspective of gender difference -- based on an empirical survey in Hubei province. *Journal of China women's university*. (2), (pp.60-65).

Sotudeh, H. Khoshian n. (2014). Gender disparities in science: The case of scientific productivity in nano science & technology during. *Scientometrics.* (98), (pp.457-472).

Sugimoto Cassidy R. Lariviere, Vincent. Ni. (2013). Chao qun Gingras Yves, Cronin, Blaise. Global gender disparities in science. *Nature.* 504, (pp.211-213).

Svein Kyvik, Mari Teigen. (1996). Child Care, Research Collaboration, and Gender Disparities in Scientific Productivity. *Science Technology Human Values.* 21(1), (pp.54-71).

Xie Yu, & Shauman K. A. (1998). Sex Disparities in Research Productivity: New Evidence about an Old Puzzle American. *Sociological Review.* 68(6), (pp.877-870).

Yuan Yuzhi. (2017). Analysis on gender difference in scientific research output of university teachers and its causes -- based on the empirical analysis of pedagogical teachers in a research university. *Exploration of higher education.* (3), (pp.5-12).

Gender gap in intellectual property rights: a case with European Union trademarks

Guillaume Roberge[1] and Matt Durning[2]

[1] *guillaume.roberge@science-metrix.com*
Science-Metrix, 1335 avenue du Mont-Royal Est, Montréal (Canada)

[2]*matt.durning@science-metrix.com*
Science-Metrix, 1335 avenue du Mont-Royal Est, Montréal (Canada)

Abstract

For decades, bibliometric statistics have chiefly relied upon scientific publications and patents to analyse research and innovation. With the recent addition of new data sources covering other relevant items (e.g., clinical guidelines, design patents, trademarks), new opportunities to enhance traditional analytical methods have arisen. Statistics on trademark activity have been gaining attention in recent years as a new intellectual property right that could be harvested to assess dimensions of innovation not covered by patent data. Gender analyses have also become mainstream over the last two decades, and although movement toward greater parity has been observed, there is still a long way to go before parity is reached in science and technology, especially in innovation, where the gender gap is more pronounced. This research-in-progress paper presents a gender analysis of trademarking activity in the European market. Comparisons with official statistics on the gender gap observed in patenting activity are used to contextualize the findings of this study because they tend to demonstrate that although parity is still far from reached in trademark activity, it is much more frequent for women to register trademarks than to apply for patents.

Introduction

The gender gap in research and technology is now a widely studied and accepted concept (Larivière, Ni, Gingras, Cronin, & Sugimoto, 2013; European Commission, 2016; Elsevier, 2017; Larivière & Sugimoto, 2017). The metrics community has studied it extensively in recent years to provide a better analytical framework for the development of policies in support of gender equality in the fields of science and technology. While most studies highlight a clear move towards better representation of women, there is still much work to be accomplished before gender parity is reached broadly, especially at later career stages (European Commission, 2016). Furthermore, while the presence of women has been increasing in scientific research, the gender gap in application and ownership of intellectual property rights such as patents is still quite pronounced, with men outnumbering women by a ratio of 10:1 in terms of patent applications within the EU28 (European Commission, 2016). Most intellectual property right (IPR) gender analysis in the past has focused on patent analyses, mostly because of limited access to data for other IPR types.

Recently, new data sources for intellectual property data have been made accessible for research. Both the European Union Intellectual Property Office (EUIPO) and the United States Patent and Trademark Office (USPTO) have made their data available to end users in recent years, making it possible to prepare gender statistics on trademark registrations. This paper presents our current work in progress on harvesting data from the EUIPO in order to analyse the gender gap in registrations of European Union trademarks (EUTM) across countries. Based on our review of the literature, only a few research papers have been published regarding gender statistics on trademark activity (Hekkllä, 2018), and we are not aware of comparable analyses based on EUIPO data. To better contextualize the gender gaps measured in trademark activity, comparisons to gender gaps measured per country in patent activity are also presented.

Methods

Data from the EUIPO were selected for this paper as we are currently in the process of preparing gender statistics on EUTM for a contract with the European Commission's Directorate-General for Research and Innovation. These trademark data are especially interesting, not only because they cover an untapped topic compared to other bibliometric metrics such as scientific papers and patents but also because of the relatively young age of this data source. Indeed, partial EUTM data have been available on an Open Data platform managed by the Office of Harmonization in the Internal Market (OHIM), now known as the EUIPO, for only a few years. However, because of anonymity issues, these data are not currently suited for full gender analyses because names of legal persons are anonymized in files from this open source. We contacted EUIPO representatives and requested and received XML files containing names of individuals in order to assign genders to applicants and prepare aggregated data according to the geographical location of trademark holders.

We assigned gender using a dictionary built as part of previous work performed by Science-Metrix during our involvement in the preparation of gender statistics on publications and patents for the latest She Figures (2016) and for the U.S. National Science Foundation (Science-Metrix, 2018). This dictionary of names was prepared with the NamSor tool, and we used it to assign a gender to holders of EU trademarks. The NamSor tool assigns names a probability of being a given gender, ranging from -1 (man) to +1 (woman), with ambiguous name combinations receiving scores closer to 0 (i.e., unisex). Overall, we retrieved about 120,000 distinct combinations of first names and last names from the EUIPO database for direct filings (i.e., filings made directly at the EUIPO office). For international registrations made through the Madrid system, data received did not include the tag for identifying the status of legal entities and names were not parsed into first and last names. Since there was no easy way to isolate individuals from organizations and to identify first names and last names, additional coding had to be performed to isolate organizations from individuals before statistics could be performed on a global scale for all trademarks registered at the EUIPO. As mentioned above, for names that were too ambiguous to be assigned a gender, a third category entitled "unisex" was created. The prominence of this category is relatively weak across most countries, accounting for 3% of all trademarks at the world level and 1% within the EU28. Only two countries stood out with notable levels of output assigned to the unisex category—China and South Korea—which both present large shares of unisex names. Findings for these countries are thus less reliable, even though results discussed later will demonstrate trends similar to other countries.

Results

A first assessment of the data is required to better understand the involvement of both men and women in terms of trademark activity. However, before delving into more details, it is important to address the legal question behind the ownership of trademarks by individuals, as opposed to ownership by a company, as it is quite a complex legal question (Pryor, 2019). There does not seem to be a consensus on which approach proves better to protect ownership of a brand. Application for a trademark must usually demonstrate that the entity requesting protection has developed the brand and has been using it, otherwise it can be opposed by other entities before final registration. Therefore, depending on whether the brand is registered by a company or an individual, the obligation to demonstrate prior use will legally lie on the shoulders of the selected legal entity, which can prove difficult if an individual registers a trademark which is used by its legally incorporated company, and not the individual itself. Conversely, ownership of a mark assigned to a company instead of an individual could lead to some problems in the

case of multiple owners owning the brand that should have been associated to an individual. These legal aspects are quite important as individual instincts about ownership differ for men and women (COFCO, 2019), which could explain parts of the patterns we will be presenting next if we suppose men are more inclined to decide on individual ownership of a brand as opposed to selecting their company, compared to women.

Data prepared demonstrate that trademark activity has been dominated by organizations, mostly from the private sector, with little past involvement from individuals, although individuals have been progressively increasing their usage of this IPR over the last two decades. Figure 1 showcases these patterns, with organizations holding about 90% of all the trademarks registered at the EUIPO, although with a share slowly but continuously decreasing in favour of more individual holders. Looking back to 1998, 96.9% of all trademarks were held by organizations, leaving 2.5% to men and a mere 0.6% for women. In contrast, by 2018, organizations held 90.2% of the newly registered trademarks, while men and women respectively held 7.0% and 2.5% of all trademarks registered that year. Consequently, the ratio of men to women, which stood at 4.6 for 1998 (i.e., 1 woman holding a trademark for each 4.6 men), experienced a sharp decrease to 2.8 by 2018. However, contribution from individuals, and especially women, remains quite limited at around 10% as of 2018.

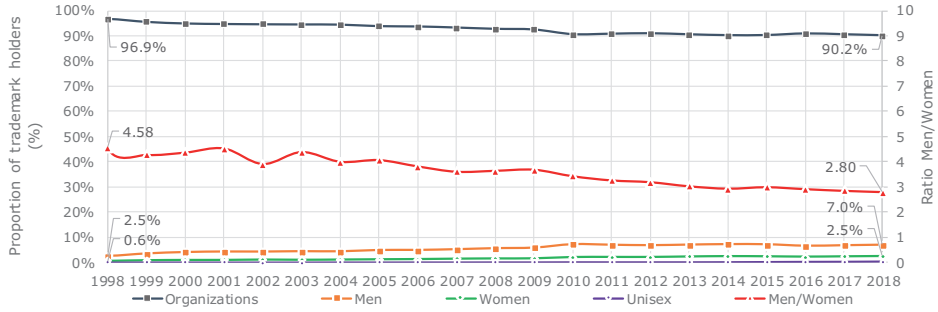


Figure 1. Proportion of EU trademark registrations held by category of holders (2000–2018)

Focusing on the 10% of all EU trademarks held by individuals makes it easier to visualize the importance of the gender gap in trademark ownership. Indeed, 70.9% of all registrations allocated to individuals in 2018 were held by men, compared to only 25.3% for women (with 3.8% falling under the unisex tag) (Figure 2).

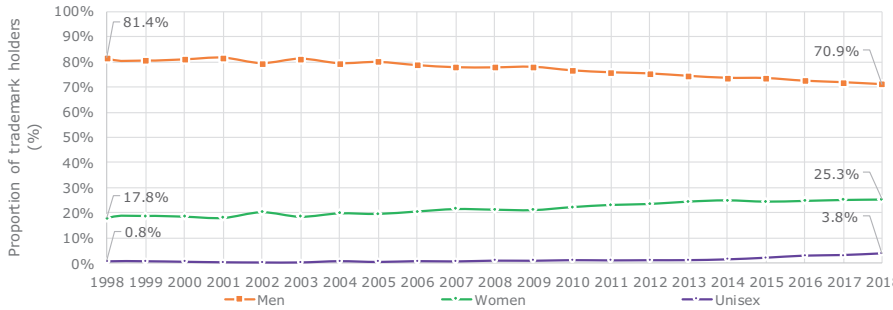


Figure 2. Proportion of EU trademark registration held amongst individuals (2000–2017)

While women's ownership of EU trademarks remains quite small, not only in terms of total ownership but also compared to that of men, female ownership has increased by 2% per year, on average, over the last 18 years (CAGR of 2%), which resulted in a net gain of 7.5 percentage points since 1998.

Even if these results in themselves are quite relevant, it is worth pointing out that without any context they are not necessarily informative on the intensity of the inequalities. For instance, is the gap observed similar to gaps measured using other metrics? To address this question, gender gap statistics on patent ownership presented in *She Figures 2015* on a country-by-country basis are presented at Table 1. While no gender gaps in patenting activity at the world level were presented in the She Figures, data at the EU28 level were available, and it can be observed that the gender gap in patenting activity is much stronger than in trademarking activity (i.e., 10 for EU patents against 2.9 for EU trademarks) (Table 1).

Table 1. Comparison of gender gaps in trademarking and patenting activities per country (2018)

Country/ Region	Percentage across all trademarks				Trademarks Individuals	Percentage trademarks (Ind.)			Gap Trademark	GAP EPO Patents (She Figures)	Ratio
	Org.	Man	Women	Unisex		Men	Women	Unisex			
World	**90%**	**7.0%**	**2.5%**	**0.4%**	**12,925**	**71%**	**25%**	**4%**	**2.8**	**N/A**	**N/A**
EU28	**88%**	**8.9%**	**3.1%**	**0.1%**	**10,439**	**74%**	**25%**	**1%**	**2.9**	**10.0**	**3.4**
Germany	84%	12.5%	3.5%	0.1%	3,032	78%	22%	0%	3.6	10.0	2.8
Spain	84%	11.5%	3.9%	0.2%	1,400	73%	25%	1%	2.9	5.0	1.7
Italy	86%	9.1%	4.4%	0.1%	1,563	67%	32%	1%	2.1	10.0	4.8
France	91%	6.6%	2.8%	0.0%	750	70%	29%	1%	2.4	10.0	4.2
China	90%	5.1%	2.6%	2.5%	1,346	51%	25%	24%	2.0	N/A	N/A
United Kingdom	93%	5.4%	1.8%	0.3%	787	73%	24%	3%	3.0	10.0	3.3
Austria	79%	16.4%	4.9%	0.0%	621	77%	23%	0%	3.3	10.0	3.0
United States	98%	1.0%	0.4%	0.0%	218	69%	28%	3%	2.5	N/A	N/A
Poland	89%	8.3%	3.0%	0.1%	354	73%	26%	1%	2.8	10.0	3.6
Belgium	88%	9.2%	3.1%	0.1%	292	74%	25%	1%	2.9	5.0	1.7
Switzerland	96%	3.0%	1.1%	0.0%	152	72%	28%	0%	2.6	10.0	3.9
Netherlands	94%	4.9%	1.4%	0.2%	277	76%	21%	3%	3.6	10.0	2.8
Czechia	88%	9.0%	3.1%	0.0%	113	74%	26%	0%	2.9	10.0	3.4
Portugal	89%	8.1%	3.1%	0.0%	169	73%	27%	0%	2.6	5.0	1.9
Hungary	73%	21.2%	5.4%	0.0%	145	80%	20%	0%	4.0	10.0	2.5
Sweden	97%	2.4%	1.1%	0.0%	105	69%	31%	0%	2.2	10.0	4.5
Romania	76%	20.5%	3.5%	0.2%	159	85%	15%	1%	5.9	5.0	0.9
Australia	98%	1.6%	0.4%	0.2%	30	73%	17%	10%	4.4	N/A	N/A
Greece	85%	10.2%	4.5%	0.0%	107	70%	30%	0%	2.3	10.0	4.4
Slovakia	87%	7.7%	5.2%	0.5%	51	58%	39%	4%	1.5	10.0	6.7
Finland	98%	1.2%	1.1%	0.0%	33	52%	46%	1%	1.1	5.0	4.4
Ireland	91%	6.3%	2.6%	0.0%	85	71%	29%	0%	2.4	10.0	4.2
Bulgaria	91%	5.9%	3.0%	0.0%	47	66%	34%	0%	2.0	10.0	5.1
Japan	100%	0.4%	0.1%	0.0%	12	75%	21%	4%	3.6	N/A	N/A
Denmark	96%	2.6%	1.0%	0.0%	60	73%	27%	1%	2.7	10.0	3.7
Canada	97%	2.0%	1.0%	0.0%	37	67%	33%	0%	2.1	N/A	N/A
South Korea	94%	1.6%	1.3%	2.7%	79	29%	23%	48%	1.2	N/A	N/A
Luxembourg	94%	4.4%	1.2%	0.0%	63	79%	21%	0%	3.8	10.0	2.6
Latvia	89%	7.0%	4.0%	0.0%	18	64%	36%	0%	1.7	2.5	1.4
Turkey	95%	3.8%	0.8%	0.0%	35	82%	18%	0%	4.6	N/A	N/A
Russia	84%	6.2%	9.7%	0.0%	69	39%	61%	0%	0.6	N/A	N/A
Ukraine	75%	18.5%	6.3%	0.0%	35	74%	26%	0%	2.9	N/A	N/A
Slovenia	88%	8.6%	3.5%	0.0%	43	71%	29%	0%	2.4	10.0	4.1
India	92%	7.2%	1.2%	0.0%	25	86%	14%	0%	6.1	N/A	N/A
Croatia	82%	11.6%	6.3%	0.0%	28	65%	35%	0%	1.8	5.0	2.7
Lithuania	90%	7.4%	3.1%	0.0%	40	71%	29%	0%	2.4	10.0	4.2
Israel	96%	3.1%	1.1%	0.0%	21	74%	26%	0%	2.8	5.0	1.8
Cyprus	95%	4.5%	0.7%	0.0%	29	86%	14%	0%	6.2	5.0	0.8
Norway	96%	3.6%	0.9%	0.0%	25	80%	20%	0%	4.0	10.0	2.5
Estonia	94%	3.9%	1.6%	0.0%	28	71%	29%	0%	2.4	3.3	1.4
South Africa	93%	5.9%	0.9%	0.0%	11	86%	14%	0%	6.3	N/A	N/A
Serbia	93%	6.4%	0.4%	0.0%	5	94%	6%	0%	15.0	N/A	N/A
Brazil	93%	5.4%	1.1%	0.0%	15	83%	17%	0%	5.0	N/A	N/A

This finding is corroborated by data on the involvement of women in patenting reported in other works (Montanari, 2018; WIPO, 2018). Impressively, this statement holds true for almost all

countries for which data were available, with an average ratio between both gender gaps of around 3 (i.e., on average, the gender gap is three times higher in patent activity compared to trademark activity across countries). Only two countries present a more pronounced gender gap for trademarks compared to patents; these are countries with low levels of trademark activity by individuals (Romania and Cyprus). All other countries for which data were presented in the She Figures present a much higher gap in patent activity.

A single country reached parity in terms of trademark ownership: Russia, where 61% of all trademarks registered in 2018 were held by women (39% for men). However, this proportion is based on fewer than 70 trademarks, so while this result is quite exciting, more data will be needed before this pattern can be confirmed. Another interesting case is South Korea, where parity was almost reached (ratio of 1.2, 29% men, 23% women), but this result is to be considered with caution because close to half of all South Korean trademark holders fell into the unisex category. The gender distribution within this category could greatly skew the result toward one gender or the other. Other countries standing out in terms of women holding registered trademarks include Finland (46%), Slovakia (39%), Latvia (36%), Croatia (35%), Bulgaria (34%), Canada (33%) and Italy (32%). At the other end of the spectrum, Serbia (6%), South Africa (14%), Cyprus (14%), India (14%), Romania (15%), Brazil (17%), Australia (17%), Turkey (18%), Norway (20%) and Hungary (20%) have the lowest percentages of women holding registered trademarks. Germany, which held the highest number of trademarks registered to individuals, fares only slightly better at 22%.

When trademarks are filed, applicants must select categories of goods and services for which their brands will be protected. The International Classification of Goods and Services (Nice classification) serves this purpose in more than 80 contracting states. Analyses at the level of these categories are quite informative as they can provide insights at a more disaggregated level and inform on differences the different sectors of the economy. Overall, there seems to be a strong alignment between what could be defined as more "traditionally women-oriented" sectors and the prevalence of women measured in trademark ownership. Indeed, the top five categories with the highest shares of women ownership cover products such as textiles and bed covers (Class 24, women representation of 38%), precious metals and jewellery (Class 14, 35%), laces, embroidery, ribbons and braids (Class 26, 34%), leather, animal skins, trunks and luggage (Class 18, 32%), and bleaching preparations and other preparations for laundry use (Class 3, 31%) (Table 2). Notably, even among these categories, women do not come close to holding the majority of trademarks, a strong reminder of their overall underrepresentation in trademark ownership. On the opposite end of the spectrum, the categories where women were least present align with areas where men are traditionally expected to be more prevalent: firearms (Class 13, 8%), machines and machine tools (Class 7, 12%), vehicles (class 12, 12%), building construction (class 37, 13%), and tobacco (class 34, 13%). Tobacco is especially interesting as although numbers tend to indicate that men do consume items covered by this category more frequently than women (16.7% of adult males and 13.6% of adult females smoke according to recent studies) (Sieminska and Jassem, 2014), the difference is not so great; yet, men are almost 7 times more like to hold a trademark related to this category than women. Focusing only on categories of services (i.e., classes 35 or above), the two categories where women were most active again mirror existing societal gender divides in terms of employment, with class 44 (medical and veterinary services, hygienic and beauty care) and class 41 (education) coming in first and second with proportions of women at 30% and 26%, respectively.

Looking at trends over time, the proportion of women within each category has remained relatively stable over the last decade, with only a handful of categories where women's presence increased by more than a few percentage points. Nevertheless, of the 45 categories, an increase between 2009–2013 and 2014–2018 was measured for 37 categories, which highlights the fact that progress, although on a small scale and at a slow pace, is occurring. Notable increases were detected for class 23 (yarns and threads, + 13.3 percentage points), class 4 (industrial oils and greases, + 9.3 percentage points), class 13 (firearms, + 6.8 percentage points), and class 8 (hand tools and implements, + 5.9 percentage points), while only one category experienced a decrease higher than 1 percentage point (tobacco, - 7.7 percentage points).

Table 2. Proportion of women holders of trademarks per category of goods and services of the Nice Classification (1998–2018)

Nice category	Label	%Women			Δ
		1998–2018	2009–2013	2014–2018	
24	Textiles and substitutes for textiles; bed covers;...	38%	40%	39%	-1.0%
14	Precious metals and their alloys; jewellery, preci...	35%	37%	37%	-0.2%
26	Lace and embroidery, ribbons and braid; buttons, h...	34%	34%	36%	1.2%
18	Leather and imitations of leather; animal skins, a...	32%	33%	35%	1.9%
3	Bleaching preparations and other substances for la...	31%	32%	33%	1.3%
27	Carpets, rugs, mats and matting, linoleum and othe...	30%	33%	35%	1.8%
44	Medical services; veterinary services; hygienic an...	30%	29%	33%	3.3%
25	Clothing, footwear, headgear...	29%	30%	31%	1.2%
16	Paper and cardboard; printed matter; bookbinding m...	28%	28%	33%	4.7%
21	Household or kitchen utensils and containers; comb...	27%	28%	28%	0.7%
41	Education; providing of training; entertainment; s...	26%	26%	29%	3.2%
4	Industrial oils and greases; lubricants; dust abso...	25%	20%	30%	9.3%
45	Legal services; security services for the physical...	25%	24%	25%	0.9%
23	Yarns and threads, for textile use...	25%	19%	32%	13.3%
20	Furniture, mirrors, picture frames; containers, no...	24%	24%	26%	1.9%
Total	Total	23%	23%	25%	1.9%
35	Advertising; business management; business adminis...	23%	23%	25%	2.3%
28	Games, toys and playthings; video game apparatus; ...	22%	22%	24%	2.3%
22	Ropes and string; nets; tents, awnings, and tarpau...	21%	23%	26%	3.1%
30	Coffee, tea, cocoa and artificial coffee; rice; ta...	21%	21%	22%	1.4%
5	Pharmaceuticals, medical and veterinary preparatio...	21%	22%	22%	0.3%
31	Raw and unprocessed agricultural, aquacultural, ho...	21%	19%	22%	3.8%
43	Services for providing food and drink; temporary a...	21%	22%	21%	-0.9%
8	Hand tools and implements (hand-operated); cutlery...	20%	17%	23%	5.9%
29	Meat, fish, poultry and game; meat extracts; prese...	19%	19%	22%	3.2%
40	Treatment of materials...	19%	21%	20%	-0.1%
2	Paints, varnishes, lacquers; preservatives against...	19%	17%	20%	2.9%
10	Surgical, medical, dental and veterinary apparatus...	19%	17%	20%	3.5%
42	Scientific and technological services and research...	18%	18%	20%	2.8%
39	Transport; packaging and storage of goods; travel ...	18%	18%	19%	1.0%
11	Apparatus for lighting, heating, steam generating,...	18%	17%	19%	1.7%
9	Scientific, nautical, surveying, photographic, cin...	17%	17%	19%	2.5%
33	Alcoholic beverages (except beers)...	16%	16%	16%	0.0%
6	Common metals and their alloys, ores; metal buildi...	16%	17%	17%	-0.3%
38	Telecommunications...	16%	15%	19%	3.4%
1	Chemicals used in industry, science and photograph...	16%	17%	17%	-0.6%
32	Beers; mineral and aerated waters and other non-al...	16%	14%	17%	3.1%
17	Unprocessed and semi-processed rubber, gutta-perch...	14%	13%	13%	0.1%
36	Insurance; financial affairs; monetary affairs; re...	14%	15%	15%	-0.2%
15	Musical instruments...	14%	13%	14%	1.7%
19	Building materials (non-metallic); non-metallic ri...	14%	14%	14%	0.3%
34	Tobacco; smokers' articles; matches...	13%	18%	10%	-7.7%
37	Building construction; repair; installation servic...	13%	14%	14%	0.3%
12	Vehicles; apparatus for locomotion by land, air or...	12%	12%	13%	1.3%
7	Machines and machine tools; motors and engines (ex...	12%	10%	14%	4.7%
13	Firearms; ammunition and projectiles; explosives; ...	8%	4%	11%	6.8%

Conclusion

After receiving EUIPO XML files containing EUTM data, we were able to prepare a set of preliminary statistics for this research-in-progress paper showcasing the huge gap between men and women regarding trademark ownership in the European market. Data demonstrate that individual ownership of trademarks has progressed over the last two decades, but the practice is still quite rare (3.1% in 1998, 9.8% by 2018). Among these, women were underrepresented by a factor of 2.8 to 1 compared to men for registrations from 2018, which was nevertheless an improvement compared to the past (4.6 to 1 in 1998, with a small but continuous increase since then). Based on these data, women had reached parity in 2018 in Russia only, but this is based on only 69 trademarks. Other notable countries regarding women's involvement in trademarking activity included Finland, Slovakia, Latvia, Croatia and Bulgaria with slightly more than a third of registered trademarks held by women, while Germany, which came first in terms of trademarks held by individuals, trailed behind with a women's representation rate of only 22%. Overall, notable higher shares for women are observed in Eastern European and Scandinavian countries compared to nations from other regions.

Comparisons with gender gaps in patent activity measured in *She Figures 2015* were also presented and the results clearly demonstrate notably weaker gender gaps (on average three times weaker) for trademark activity, and this across almost all countries. One hypothesis for this finding is that the traditional barriers for women in industries, where most of the patentable research and innovation are being performed, are not as strong in the realms where trademarks are relevant (e.g., in small enterprises providing services). Salient findings at the level of Nice categories of goods and services demonstrated that some categories have a larger share of their trademarks held by individuals attributed to women, but none of the 45 categories has firmly entered the parity zone yet (i.e., at least 40% women). Further research will be needed to understand the weaker gender gap measured in trademarks compared to patents.

Another path of investigation for the future is to prepare similar country statistics for the American market using data from the USPTO to detect any notable differences between both markets. Country-to-country comparisons would be performed to detect a possible effect of the internationalization of trademarks on gender-based ownership. One hypothesis, somewhat analogous to the weaker presence of women on international collaboration partnerships, is that women may tend to less frequently register trademarks in foreign markets compared to men. Consequently, the proportion of women amongst individual holders in European nations should decrease when measured in the American market, while proportions of US women (and other North American nations such as Canada and Mexico) holding trademarks should decrease going from the USPTO to the EUIPO.

Acknowledgments

We thank the European Union Intellectual Property Office (EUIPO) for access to XML data files on European Union trademarks through their open data platform and licensing agreement for access to non-anonymized data. A special thanks is owed to Pablo Garcia Ortega, a representative from the EUIPO, for his support with these data, as well as for his kind support during our data validation process.

References

Larivière, V., Ni, C., Gingras, Y., Cronin, B., & Sugimoto, C. (2013). Bibliometrics: Global gender disparities in science. *Nature*, 504 (7479), 211–213.

European Commission. (2016). *She Figures 2015*. Luxembourg: Publications Office of the European Union.

Elsevier. (2017). *Gender in the global research landscape*. Retrieved from https://www.elsevier.com/__data/assets/pdf_file/0008/265661/ElsevierGenderReport_final_for-web.pdf

Larivière, V., & Sugimoto, C. (2017). The end of gender disparities in science? If only it were true… CWTS. Retrieved online on January 15, 2019 at https://www.cwts.nl/blog?article=n-q2z294

Hekkllä, J. (2018). IPR gender gaps: a first look at utility model, design right and trademark filings. *Scientometrics*, 1–15.

Pryor, J (2019). Should I own my trademark or should my company? Quick of the mark. Retrieved online on May 30, 2019 at http://www.http://www.quickoffthemark.com.au/blog/should-i-own-my-trademark-or-should-my-company/

COFCO (2019). Men owned businesses vs. women owned businesses… Are there differences? COFCO. Retrieved online on May 30, 2019 at http://www.http://www.cofcogroup.com/men-owned-businesses-vs-women-owned-businesses-are-there-differences/

Montanari, L. (2018). How IP rights empower women. *Forbes*. Retrieved from https://www.forbes.com/sites/lorenzomontanari/2018/04/26/how-ip-rights-empower-women/#6f09cbce6e73

Sieminska, A., Jassem, E (2014). The many faces of tobacco use among women. Medical Science Monitor: *International Medical Journal of Experimental and Clinical Research*, 153–162.

Science-Metrix. (2018). *Development of bibliometric indicators to measure women's contribution to scientific publications.* Retrieved from http://www.science-metrix.com/sites/default/files/science-metrix/publications/science-metrix_bibliometric_indicators_womens_contribution_to_science_report.pdf

WIPO. (2018). World Intellectual Property Indicators 2018. WIPO. Retrieved from https://www.wipo.int/edocs/pubdocs/en/wipo_pub_941_2018.pdf

Exploring the historical roots of Mesenchymal stem cell research using reference publication year spectroscopy

Adil El Aichouchi[1] and Philippe Gorry[2]

[1] *adil.el-aichouchi@u-bordeaux.fr*
Centre Emile Durkheim, UMR CNRS 5116, Dpt. of Humanities and Social Science, University of Bordeaux, 3 ter Place de la Victoire 33076 Bordeaux (France)

[2] *philippe.gorry@u-bordeaux.fr*
GREThA UMR CNRS 5113, Dpt. of Humanities and Social Science, University of Bordeaux, Av. Leon Duguit, 33608, Pessac (France)

Abstract

Mesenchymal stem cells have been in the focus of research in the emerging field of regenerative medicine. It is also one of the most confusing and controversial. This paper applies Reference publication Year Spectroscopy (RPYS) in order to explore the historical roots of mesenchymal stem cell research by identifying the key historical publications that initiated the field. Analysis of top 10 cited references identified by the RPYS methodology revealed four papers that suffered from Delayed Recognition in terms of citations (DRs). Interestingly, all the DRs identified belong to a Russian scientist, Alexander J. Friedenstein, who is widely regarded as the discoverer of mesenchymal stem cells (MSCs). The paper explains the reasons behind the delay in recognition of Friedenstein's work using historical and sociological analysis.

Introduction

Mesenchymal stem cells

For years mesenchymal stem cells (MSCs) have been in the focus of research in the emerging field of regenerative medicine. In their analysis of the global stem cell research trend during the previous decade, Li et al. (2009) identified MSCs one of the main orientations of all the stem cell research in the 21st century. Subsequently, in their scientometric analyses of the regenerative medicine landscape, Chen et al. (2014; 2012) found that mesenchymal stem cell research was one of the main emerging research lines in the field. MSCs are multipotent stromal cells that can differentiate into a variety of cell types including osteoblasts (bone cells), chondrocytes (cartilage cells), myocytes (muscle cells), and adipocytes (fat cells). MSCs are seen as excellent candidates for therapeutic use as cellular therapies that have the potential to revolutionize the current pharmaceutical landscape (Gao et al., 2016). As of August 2018, a search using the keyword "Mesenchymal stem cells" in the ClinicalTrials database (http://clinicaltrials.gov) identified 866 clinical trials at different stages and in a variety of diseases. However, research on MSCs is permeated by a much confusion and controversy (Ankrum, Ong, & Karp, 2014; Bianco & Gehron Robey, 2000; Gao et al., 2016; Sacchetti et al., 2016; Sharpe, 2016; Wagner & Ho, 2007). As Sacchetti and colleagues recently wrote: "the anatomical identity of MSCs, their phenotype, distribution in different tissues, lineage, physiological functions, and biological properties represent one of the most confusing areas in stem cell biology" (Sacchetti et al., 2016). MSCs are thought to have inconsistent protocols, varying dosages and differing transfusion patterns (Gao et al., 2016). In addition, the definition of mesenchymal stem cells is based solely on their *in vitro* characteristics (Sharpe, 2016), and it is unclear if the MSC phenotype exists *in vivo* (Ankrum et al., 2014; Wang, Chen, Cao, & Shi, 2014). Consequently, all MSCs are now considered to be the same, regardless of their tissue of origin, to the extent that MSCs from widely different tissues are often considered equivalent in clinical applications (Sharpe, 2016). Also, the nomenclature of MSCs remains

contentious, and the use of the term "Mesenchymal stem cell", coined in the early 1990s by Caplan et al. (1991) is contested by many (Bianco & Gehron Robey, 2000; Bianco, Robey, & Simmons, 2008; da Silva Meirelles, Caplan, & Nardi, 2008; Phinney, 2002) mainly because of the unclear *in vivo* identity of MSCs. This state of affairs have also pushed the International Society for Cellular Therapy to issue a position paper to clarify the nomenclature of MSCs (Dominici et al., 2006). This makes MSC research a highly dynamic and controversial area in stem cell research.

Reference publication year spectroscopy

Novel research builds on previous ideas and discussions in the scientific literature (Merton, 1965; Ziman, 2000). Moreover, the correlation between the importance of a publication and number of citations it has is the premise of a normative theory of citation (Bornmann, de Moya Anegón, & Leydesdorff, 2010; Merton, 1965). Based on these assumptions, it is possible to use citation data to gain historical insight on the importance of previous publications to a given research field. "Reference Publication Year Spectroscopy" (RPYS) is a quantitative method that reveals the most important historical publications of a specific research field based on the analysis of the publication years of the references cited within the relevant literature (Marx, Bornmann, Barth, & Leydesdorff, 2014). RPYS exploits the fact that the analysis of the publication years of the references cited by the papers in a specific research field reveals pronounced peaks in the distribution of the reference publication years (Marx, Haunschild, Thor, & Bornmann, 2017). This means that some years stand out as more significant than the others, and these peaks are usually based on single early publications of significant importance to the field.

RPYS has been used to identify the historical roots of climate change (Marx et al., 2017), graphene research (Marx et al., 2014), the development and use of the global positioning system (Comins & Hussey, 2015) as well as philosophy of science (Wray & Bornmann, 2015). In addition to identifying the historical references of a given field, it is useful to analyze the citation distributions of the identified CRs to gain insights into their impact on the field. In general, citations accumulate in some common patterns over time. Costas et al. (2010) distinguish three types of citation histories: (1) Normal: where documents reach a maximum in, say, 3 to 4 years after publication and then decay exponentially; (2) Flashes in the pan: where documents receive citations immediately after their publication, but are not cited in the long term; (3) Delayed: where documents receive the main part of their citations later than normal documents. An extreme case of the latter category is what used to be called in scientometrics "Sleeping Beauties" (Van Raan, 2004). We will call these papers Delayed Recognition papers (DRs) to use a gender-neutral name (Sugimoto & Mostafa, 2018) and to avoid confusion with other uses of the same metaphor in other disciplines. DRs are papers that receive very little or no citations shortly after publication before finally getting cited after few years (Are "awakened") (Garfield, 1989b, 1989a, 1990; Van Calster, 2012; Van Raan, 2004). Given the age of the top CRs identified by RPYS, and the fact that they are still of current value to the field, it is important to explore their citation histories.

The present paper aims at identifying the intellectual roots of mesenchymal stem cell research using the RPYS methodology on a dataset of mesenchymal stem cell publications and see among those publications which ones have the highest impact in this research field by analyzing their citation distributions.

Materials and Methods

The analyses are based on a publication dataset from the Web of Science Core Collection produced by Thomson Reuters (Philadelphia, USA). The dataset is the result of the following query: ("CFU F" OR "CFU-Fibroblast*" OR "colony forming unit fibroblast*" OR "colony

forming unit* fibroblast*" OR "bone-marrow stromal stem cell*" OR "bone marrow fibroblast*" OR "bone stem cell*" OR "fibroblast* colony forming cell*" OR "Fibroblast* colony forming unit*" OR "fibroblast* CFU*" OR "fibroblast* precursor*" OR "marrow-isolated adult multilineage inducible cell*" OR "marrow stromal cell*" OR "marrow stromal stem cell*" OR "Marrow stromal fibroblast"* OR "mesenchymal stem cell*" OR "mesenchymal stromal cell*" OR "mesenchymal progenitor cell*" OR "mesenchymal stromal cell*" OR "multipotent mesenchymal stromal cell*" OR "multipotent adult progenitor cell*" OR "Multipotent stromal cell*" OR "multipotent adult progenitor cell*" OR "Medicinal signaling cell*" OR "osteogenic stem cell*" OR "stromal stem cell*" OR "skeletal stem cell*" OR "very small embryonic-like cell*" OR "Unrestricted somatic stem cell*"). The query is the result of an extensive review of the MSC literature as well as semi-structured interviews with 5 expert researchers in the MSC field. It aims to contain all the different ways in which scientists describe MSCs, and thus captures the complex and confusing aspect of its nomenclature. The query yields a dataset of n=57,718 publications (articles and reviews only). Once the dataset with complete metadata was extracted, RPYS analysis was applied on the dataset using a custom Python script. Metaknowledge, a Python package for computational research in information science, network analysis, and science of science was used to extract the cited references of each publication in the dataset and organize them in the appropriate data structure for further analysis (McLevey & McIlroy-Young, 2017). The list of raw references (n=3,289,601) was then cleaned and transformed into strings with a unified structure: "name of the first author, publication year, abbreviated journal name, volume, first page, doi" e.g: "caplan ai, 1991, j orthop res, v9, p641, 10.1002/jor.1100090504". This list of raw references is the aggregation of all the CRs of each article in the dataset. At this step of the process, if an article was cited 100 times in the dataset, it will be represented 100 times in the list of raw references. The next step consists of merging the instances of each unique CR into one data point, and calculating its frequency (i.e. times cited by the dataset). This merging procedure creates a smaller list of unique references (n = 979,587). This was done by applying inbuilt Python functions as well as the Jaro-Winkler similarity measure with a threshold of 0.95 (Winkler, 1990). Further calculations on the distribution of the cited references were done in Python.

Co-author and citation networks were created using VOSviewer, an open source software for bibliometric mapping (van Eck & Waltman, 2010), and Gephi, an open-source software for network analysis (Bastian, Heymann, & Jacomy, 2009). Data used to create the coauthor and citation networks were extracted from Web of Science as well.

To identify DR papers in a dataset, we calculate the "Beauty coefficient" (B), a parameter-free index introduced by Ke et al. (Ke, Ferrara, Radicchi, & Flammini, 2015). "B" quantifies the extent to which a paper could be considered a DR paper by adding up differentials between the citation curve of the publication and a reference line calculated between the year of publication and the year of maximum citations. Applying their criteria to their database (n=22,379,244), Ke et al. found that the top 1,000 DR papers in their database correspond to papers with B ≥ 317.93.

Using Ke et al's criteria (Ke et al., 2015), the "Beauty coefficient" B was calculated for the top 10 cited references in the dataset that were published before 1991.

Results

The dataset contains n=57, 718 publications with the year with maximum publications being 2017 (n=7,465) (Figure 1). This reflects the dynamism over the last twenty years of MSC research in particular and stem cell research and regenerative medicine in general. The number of CRs per publication averaged 59.44. The distribution of CRs starts in 1543 (Vesalius a, 1543, humani corporis fabr) and reaches its maximum in 2008 (n=218,772) (Figure 2, blue box). For clarity, the distribution of CRs between 1543 and 1899 is not shown in Figure 2.

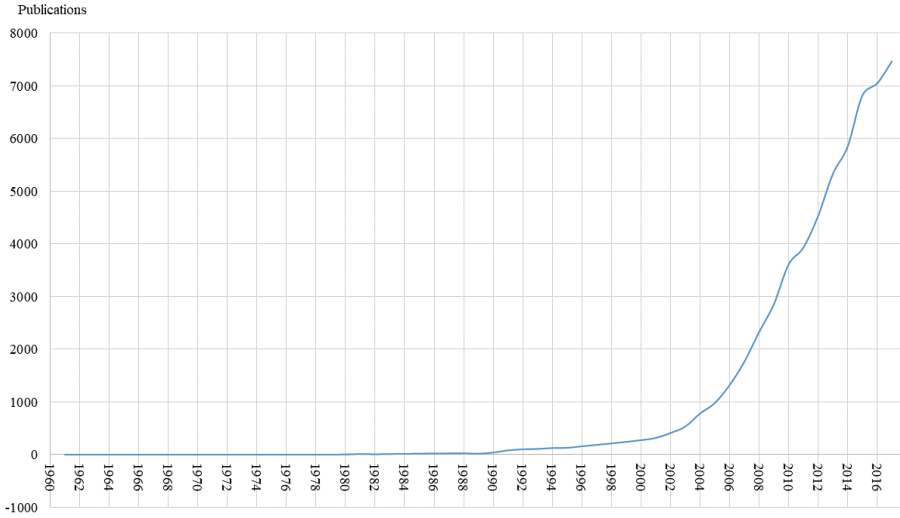


Fig 1. Distribution of publications in the MSC dataset

The horizontal black line in Figure 2 ("Deviation from the 5-year median") shows the deviation of the number of CRs in each year from the median for the number of CRs in the two previous, the current, and the two following years (Y-2; Y-1; Y; Y+1; Y+2). The deviation from the 5-year median plays the role of a curve smoother and accentuates the peaks in the data better than absolute numbers given that each year is compared with its adjacent years (Thor, Marx, Leydesdorff, & Bornmann, 2016). The most noticeable peaks in Figure 2 happen between 1960 and 1980. A much later peak occurs between 2007 and 2008, and is explained by the fact that these two years have the two highest numbers of cited references (n=217,454 and n=218,772 respectively) (Figure 2, vertical bars).

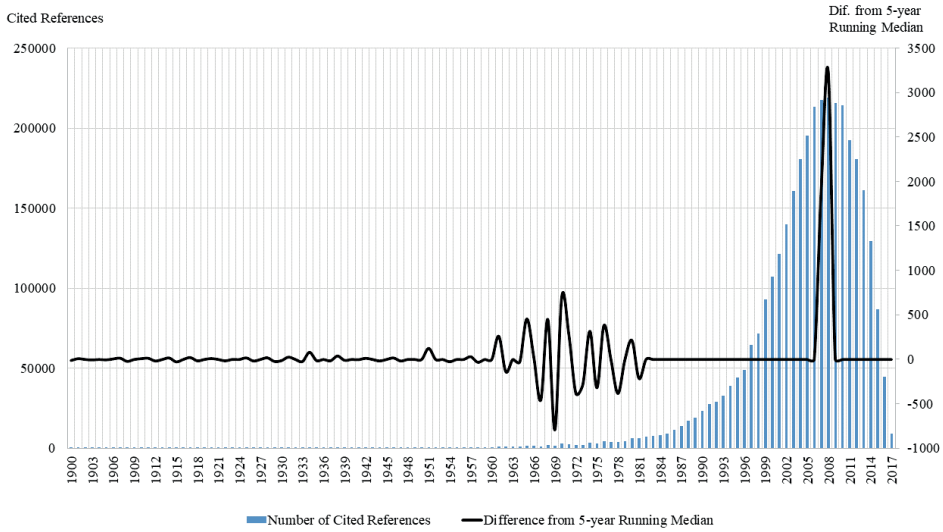


Fig 2. Reference Publication Year Spectroscopy of the MSC dataset.

To identify the CRs responsible for these peaks, we sorted the list of unique CRs published before 1991 by their frequency of citations (Table 1). CRs #1, #2, #3, #5, #6, #7, and #9 in Table 1 are all articles coauthored by Russian scientist Alexander Jakovlevich Friedenstein (1924-1997) and his colleagues and collaborators at the Gamaleya Institute for Epidemiology and Microbiology in Moscow USSR, with the unique exception of his collaboration with Maureen Owen from the MRC Bone Research Laboratory in Oxford University in the late 1980s (Table 1, #5). In this bundle of highly related publications, Friedenstein and colleagues described fibroblastoid cells that were obtained from bone marrow showing colony formation and *in vitro* as well as *in vivo* osteogenic differentiation potential. They established that cells that are adherent, clonogenic, nonphagocytic, and fibroblastic in habit (Colony forming units-fibroblastic or CFU-F) can be isolated from the bone marrow stroma of post-natal organisms. They can also give rise, under appropriate experimental conditions to a broad spectrum of fully differentiated connective tissues including cartilage, bone, adipose tissue, fibrous tissue, and myelosupportive stroma (Bianco & Gehron Robey, 2000). They utilized *in vitro* culture and transplantation in laboratory animals, either in closed systems like diffusion chambers (e.g. in #1) or open systems, subcutaneously (e.g. in #7) or under the renal capsule (e.g. in #3 and #5), to characterize cells that compose the physical stroma of bone marrow (Bianco, Riminucci, Gronthos, & Robey, 2001). In CR #10, Hugo Castromalaspina and colleagues extended their work by characterizing human bone marrow CFU-Fs and their progeny.

In CR #4 Maniatopoulos and colleagues developed conditions in which cells obtained from adult bone marrow could form mineralized bone modules. The method described in the paper is widely used for studies of mineralization and the identification of osteogenic progenitor cells (McCulloch, 1999).

CR #8 is the result of pioneering work by the American orthopedic surgeon Marshall R. Urist, where he showed that new bone formation could be induced by implanting the demineralized bone matrix under the skin or into the muscle of animals. He also discovered bone morphogenetic proteins that induce the formation of bone and cartilage.

Discussion

Many reasons have been explored in the literature as to why the phenomenon of delayed recognition in terms of citations occurs and what mechanisms are in play in the "rediscovery" or the "awakening" of DR papers. Some publications "sleep" for a long time because they contain scientific claims that steer controversy in a given field, or because they are at the boundary between two disciplines (Gorry & Ragouet, 2016). This "sleeping" period only ends when the controversy is closed, and when consensus is reached. Other publications don't get enough citations in the beginning because they contain unproven hypotheses, and the citational "awakening" occurs when an experimental proof of the hypothesis is presented (El Aichouchi & Gorry, 2018a). In some applied fields, an industrial application is needed to make a forgotten publication relevant again (El Aichouchi & Gorry, 2018b). In some cases, the DR papers were simply not available to a wide enough scientific community (Garfield, 1980; Gorry & Ragouet, 2016; Hook, 2002), either due to language barriers or lack of visibility due to its publication in unknown journals.

Table 1. Top 10 cited references in the dataset before 1991. The awakening year and the means are only calculated when relevant (i.e. the value of B is high)

	Reference	Year	Total citations	Citations from the dataset	B	Awakening year (A)
#1	Friedenstein, AJ, Chailakhjan, RK and Lalykina, KS, The development of fibroblast colonies in monolayer cultures of Guinea-pig bone marrow and spleen cells. Cell Proliferation, 3: 393-403.	1970	1,320	945	288.04	2001
#2	Friedenstein, AJ, Gorskaja, UF, Kulagina, NN, Fibroblast precursors in normal and irradiated mouse hematopoietic organs. Experimental Hematology, 4: 267-274	1976	1,010	793	440.20	2002
#3	Friedenstein, AJ, Petrakova, KV, Kurolesova, AI, *et al.*, Heterotopic transplants of bone marrow – Analysis of precursor cells for osteogenic and hematopoietic tissues. Transplantation, 6: 230-247	1968	1,096	735	264.51	2001
#4	Maniatopoulos, C, Sodek, J, Melcher AH, Bone formation in vitro by stromal cells obtained from bone-marrow of young-adult rats. Cell Tissue Research, 254:317-330	1988	1124	633	0.3	-
#5	Owen, M, Friedenstein, AJ, Stromal stem cells – marrow-derived osteogenic precursors. CIBA Foundation Symposia, 136:42-60	1988	789	602	9.87	-
#6	Friedenstein, AJ, Chailakhyan, RK, Latsinik, NV, *et al.*, Stromal cells responsible for transferring microenvironment of hematopoietic tissues – cloning *in vitro* and retransplantation *in vivo*. Transplantation, 17: 331-340	1974	915	586	74.19	-
#7	Friedenstein, AJ, Piatetzk, II, Petrakova, KV, Osteogenesis in transplants of bone marrow cells. Journal of Embryology and Experimental Morphology, 16:381 - ,	1966	836	570	529.73	2001
#8	Urist, MR, Bone formation by autoinduction. Science. 150:893	1965		552	70.78	-
#9	Friedenstein, AJ, Chailakhyan, RK, Gerasimov, UV, Bone marrow osteogenic stem cells – *in vitro* cultivation and transplantation in diffusion chambers. Cell and Tissue Kinetics, 20: 263-272	1987	634	485	11.52	-
#10	Castromalaspina, H, Gay, RE, Resnick, G, *et al.*, Characterization of human bone marrow fibroblast colony-forming cells (CFU-F) and their progeny. Blood, 56:289-301	1980	665	482	19.40	-

Looking at the B score of the CRs in Table 1, as well as their citation distributions in Figure 3, it is clear that CRs #1, #2, #3, and #7 are outstanding DR papers. The awakening year of these papers ranges between 2001 and 2002 (Awakening Year (A); Table 1).

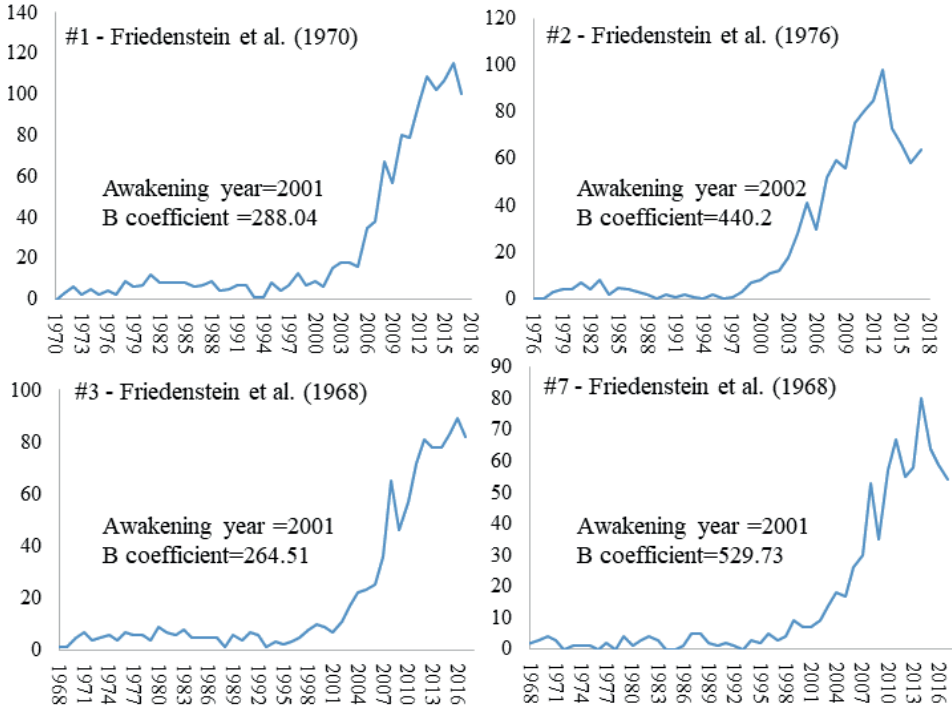


Fig 3. Citation distribution of publications #1, #2, #3, and #7

In Friedenstein's case, the size of his social network played a huge role in the limited diffusion of his scientific ideas. Although it is true that scientific publications are one of the cornerstones of scholarly communication, publishing in peer-reviewed journals is not always enough for the diffusion of novel scientific ideas. Social networks, attending scientific meetings, and peer discussions are extremely important in that process because they provide occasions for the diffusion of additional complementary knowledge that is tacit and thus impossible to codify (Agrawal, Cockburn, & McHale, 2006; Ganguli, 2015). In addition, the accumulation of human capital occurs when people are able to learn and interact with other individuals who have a close geographic an social proximity (Agrawal et al., 2006; Glaeser, 1999).

This was not Friedenstein's case. He and his colleagues suffered from "long years of now infamous stagnation, short supply and severely restricted personal contacts" (Kuznetsov & Robey, 2000). The Wilsede meeting in 1988 was only "the second time that he [Friedenstein] has been outside the USSR at a scientific meeting", according to Michael A.S. Moore, who presented Friedenstein at the meeting (Friedenstein, 1988).

This is not to say that nobody was aware of Friedenstein and his collaborators' work in the 1970s. In 1977, in a forward to a translation of a book by Elena Luria, Friedenstein's wife and one of his long-term collaborators (Figure 4), Robert Auerbach wrote that "while most of us have been well aware of the contributions to the literature made by research work carried out by Drs. Fridenshtein and Luriya, investigators such as Kolesnikova, Chailakhyan, Chertkov, Latsinik, and Keilis-Borok, whose names appear as senior authors in studies of obvious significance in the field, are virtually ignored outside the USSR because of the restriction of their published work to the Soviet literature" (Luriya, 1977).

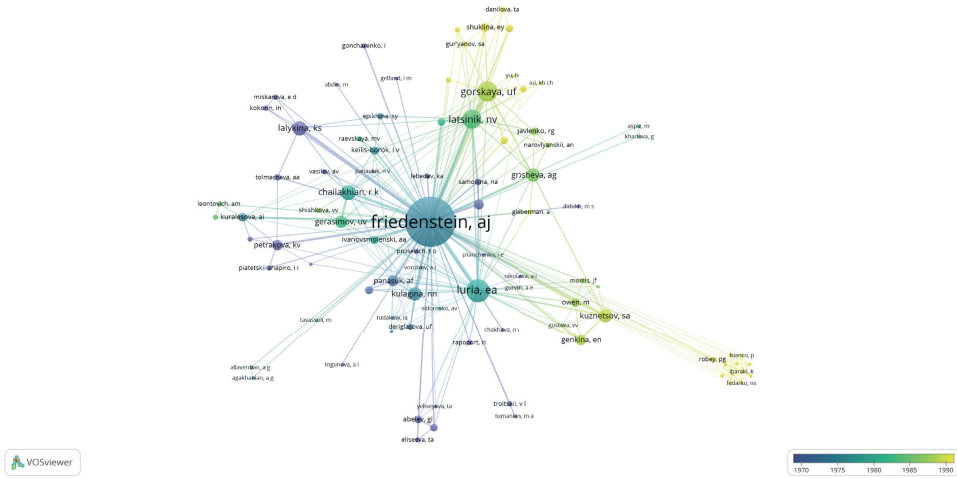


Fig 5. Friedenstein's co-authorship network. The color of the node reflects the average

The names cited by Auerbach are all close collaborators of Friedenstein. The unavailability of their work, and their collective isolation at the international level is a huge obstacle for the spreading of their theories, protocols, assays, and skills.

The first time that Friedenstein had collaborated with a scientist from outside the Eastern-bloc was when he co-authored a paper in 1983 with Mehdi Tavassoli, a scientist from the Veterans Administration Hospital, University of Mississippi Medical Center, Jackson (M. T) (Tavassoli & Friedenstein, 1983) (Figure 4). The rise to power of Gorbachev in 1985 and the beginning of the period of "perestroika" provided a shift toward less isolationism for Soviet scientists (Vasiliev, 2008). Around that time, Friedenstein started a very fruitful collaboration with British scientist Maureen Owen, from the MRC Bone Research Laboratory, Nuffield Orthopaedic Center, Oxford, UK. This collaboration resulted in the publication of their 1988 classic paper (CR #5). In the 1990s, the Soviet Union had ended, which provided Russian scientists even more opportunities to migrate to Western countries and to attend scientific meetings. Friedenstein benefited from this situation too. From 1991 until 1994, and thanks to Pamela Gehron Robey, Friedenstein was a visiting researcher and a "special volunteer" at the Bone Research Branch in the National Institute of Dental Research (NIH, Bethesda, Maryland) (Robey, 1992, 1993, 1994). This occasion gave him the possibility, although late in his life, to extend his social network and to share his knowledge more effectively. Friedenstein's case exemplifies perfectly the finding that ideas in high impact Soviet papers and papers previously accessible to US scientists were the most likely to "spill over" to natives (Ganguli, 2015).

Another factor that contributed to Friedenstein's late wakening of his DR papers is the ethical issues surrounding embryonic stem cells. Despite their pluripotency, the moral controversies surrounding their use in therapy and clinical application pushed the stem cell research community to examine other types of progenitor cells (Tuan, Boland, & Tuli, 2002). Especially effective was George W. Bush's policy in 2001 that limited the number of stem cell lines that could be used for research (Robertson, 2010).

Conclusion

The examination of the "Mesenchymal stem cell" concept via RPYS identified a set of publications by Russian scientist Alexander Friedenstein that proved foundational to the

emergence and development of an MSC research program that is currently a highly dynamic field and a highly controversial one.

The analysis of the citation distribution of the CRs identified by the RPYS reveals a bundle of highly cited and highly related papers by Friedenstein, Maureen E. Owen, Hugo Castromalaspina, as well as Marshall Urist that were foundational for MSC research later on. A subset of these papers can be described as papers with delayed recognition in terms of citations. The reasons behind this delay can be attributed to the geopolitical situation during the Cold War that hampered the diffusion of scientific ideas coming from the Soviet Union through lack of communication and collaboration with Western scientists. The end of the Soviet Union gave Friedenstein and colleagues the opportunity to finally extend their social network and promote his ideas better. In addition, the ethical controversies surrounding embryonic stem cells pushed the stem cell community to intensify their search for alternatives, which put MSCs and thus Friedenstein's research in the spotlight again.

Acknowledgments

The authors would like to thank Dr. Zoran Ivanovic for his valuable insights and comments as well Helene Boeuf, …. Adil El Aichouchi is supported by an interdisciplinary PhD fellowship from the University of Bordeaux.

References

Agrawal, A., Cockburn, I., & McHale, J. (2006). Gone but not forgotten: knowledge flows, labor mobility, and enduring social relationships. *Journal of Economic Geography*, *6*(5), 571–591. Retrieved from http://dx.doi.org/10.1093/jeg/lbl016

Ankrum, J. A., Ong, J. F., & Karp, J. M. (2014). Mesenchymal stem cells: immune evasive, not immune privileged. *Nature Biotechnology*, *32*, 252. Retrieved from http://dx.doi.org/10.1038/nbt.2816

Bastian, M., Heymann, S., & Jacomy, M. (2009). Gephi: an open source software for exploring and manipulating networks. *Icwsm*, *8*, 361–362. https://doi.org/10.13140/2.1.1341.1520

Bianco, P., & Gehron Robey, P. (2000). Marrow stromal stem cells. *The Journal of Clinical Investigation*, *105*(12), 1663–1668. https://doi.org/10.1172/JCI10413

Bianco, P., Riminucci, M., Gronthos, S., & Robey, P. G. (2001). Bone Marrow Stromal Stem Cells: Nature, Biology, and Potential Applications. *Stem Cells*, *19*(3), 180–192. https://doi.org/10.1634/stemcells.19-3-180

Bianco, P., Robey, P. G., & Simmons, P. J. (2008). Mesenchymal Stem Cells: Revisiting History, Concepts, and Assays. *Cell Stem Cell*, *2*(4), 313–319. https://doi.org/http://dx.doi.org/10.1016/j.stem.2008.03.002

Bornmann, L., de Moya Anegón, F., & Leydesdorff, L. (2010). Do Scientific Advancements Lean on the Shoulders of Giants? A Bibliometric Investigation of the Ortega Hypothesis. *PLOS ONE*, *5*(10), e13327. Retrieved from https://doi.org/10.1371/journal.pone.0013327

Caplan, A. I. (1991). Mesenchymal stem cells. *Journal of Orthopaedic Research*, *9*(5), 641–650. https://doi.org/10.1002/jor.1100090504

Chen, C., Hu, Z., Liu, S., & Tseng, H. (2012). Emerging trends in regenerative medicine: a scientometric analysis in CiteSpace. *Expert Opinion on Biological Therapy*, *12*(5), 593–608. https://doi.org/10.1517/14712598.2012.674507

Chen, C., Dubin, R., & Kim, M. C. (2014). Emerging trends and new developments in regenerative medicine: a scientometric update (2000 – 2014). *Expert Opinion on Biological Therapy*, *14*(9), 1295–1317. https://doi.org/10.1517/14712598.2014.920813

Comins, J. A., & Hussey, T. W. (2015). Detecting seminal research contributions to the development and use of the global positioning system by reference publication year spectroscopy. *Scientometrics*, *104*(2), 575–580. https://doi.org/10.1007/s11192-015-1598-2

Costas, R., van Leeuwen, T. N., & van Raan, A. F. J. (2010). Is scientific literature subject to a 'Sell-By-Date'? A general methodology to analyze the 'durability'of scientific documents. *Journal of the American Society for Information Science and Technology*, *61*(2), 329–339. https://doi.org/10.1002/asi.21244

da Silva Meirelles, L., Caplan, A. I., & Nardi, N. B. (2008). In Search of the In Vivo Identity of Mesenchymal Stem Cells. *STEM CELLS*, *26*(9), 2287–2299. https://doi.org/10.1634/stemcells.2007-1122

Dominici, M., Le Blanc, K., Mueller, I., Slaper-Cortenbach, I., Marini, F. C., Krause, D. S., Deans, R., Keating, A., Prockop, D. J., Horwitz, E. M. (2006). Minimal criteria for defining multipotent mesenchymal stromal cells. The International Society for Cellular Therapy position statement. *Cytotherapy*, *8*(4), 315–317. https://doi.org/http://dx.doi.org/10.1080/14653240600855905

El Aichouchi, A., & Gorry, P. (2018a). Delayed recognition of Judah Folkman's hypothesis on tumor angiogenesis: when a Prince awakens a Sleeping Beauty by self-citation. *Scientometrics*. https://doi.org/10.1007/s11192-018-2752-4

El Aichouchi, A., & Gorry, P. (2018b). Paul Hagenmuller's contribution to solid state chemistry: A scientometric analysis. *Journal of Solid State Chemistry*, *262*, 156–163. https://doi.org/https://doi.org/10.1016/j.jssc.2018.02.003

Friedenstein, A. J. (1988). On stromal-hematopoietic interrelationships. Maximov's ideas and modern models. Wilsede. https://doi.org/10.3205/wsc-2009-en-000041.01

Ganguli, I. (2015). Immigration and Ideas: What Did Russian Scientists "Bring" to the United States? *Journal of Labor Economics*, *33*(S1), S257–S288. https://doi.org/10.1086/679741

Gao, F., Chiu, S. M., Motan, D. A. L., Zhang, Z., Chen, L., Ji, H.-L., Tse, H. F., fu, Q. L., Lian, Q. (2016). Mesenchymal stem cells and immunomodulation: current status and future prospects. *Cell Death and Disease*, *7*, e2062. Retrieved from http://dx.doi.org/10.1038/cddis.2015.327

Garfield, E. (1980). Premature discovery or delayed recognition-Why? *Current Contents*, (21), 5–10.

Garfield, E. (1989a). Delayed recognition in scientific discovery-citation frequency-analysis aids the search for case-histories. *Current Contents*, *23*, 3–9.

Garfield, E. (1989b). More delayed recognition. 1. Examples from the genetics of color-blindness, the entropy of short-term-memory, phosphoinositides, and polymer rheology. *Current Contents*, *38*, 3–8.

Garfield, E. (1990). More delayed recognition. 2. From inhibin to scanning electron microscopy. *Current Contents*, *9*, 3–9.

Glaeser, E. L. (1999). Learning in Cities. *Journal of Urban Economics*, *46*(2), 254–277. https://doi.org/https://doi.org/10.1006/juec.1998.2121

Gorry, P., & Ragouet, P. (2016). "Sleeping beauty" and her restless sleep: Charles Dotter and the birth of interventional radiology. *Scientometrics*, *107*(2), 773–784. https://doi.org/10.1007/s11192-016-1859-8

Hook, E. B. (2002). A background to prematurity and resistance to "discovery." *Prematurity in Scientific Discovery: On Resistance and Neglect*, 3–21.

Ke, Q., Ferrara, E., Radicchi, F., & Flammini, A. (2015). Defining and identifying Sleeping Beauties in science. *Proceedings of the National Academy of Sciences* , *112*(24), 7426–7431. https://doi.org/10.1073/pnas.1424329112

Kuznetsov, S. A., & Robey, P. G. (2000). A look at the history of bone marrow stromal cells. *Graft*, *3*(6), 278–283.

Li, L.-L., Ding, G., Feng, N., Wang, M.-H., & Ho, Y.-S. (2009). Global stem cell research trend: Bibliometric analysis as a tool for mapping of trends from 1991 to 2006. *Scientometrics*, *80*(1), 39–58. https://doi.org/10.1007/s11192-008-1939-5

Luriya, E. A. (1977). *Hematopoietic and lymphoid tissue in culture*. Springer Science & Business Media. https://doi.org/10.1007/978-1-4684-1629-9

Marx, W., Bornmann, L., Barth, A., & Leydesdorff, L. (2014). Detecting the historical roots of research fields by reference publication year spectroscopy (RPYS). *Journal of the Association for Information Science and Technology*, *65*(4), 751–764. https://doi.org/10.1002/asi.23089

Marx, W., Haunschild, R., Thor, A., & Bornmann, L. (2017). Which early works are cited most frequently in climate change research literature? A bibliometric approach based on Reference Publication Year Spectroscopy. *Scientometrics*, *110*(1), 335–353. https://doi.org/10.1007/s11192-016-2177-x

McCulloch, C. A. (1999). Tony Melcher's Contributions to the Regeneration of the Periodontium. *Journal of Dental Research*, *78*(7), 1292–1297. https://doi.org/10.1177/00220345990780070201

McLevey, J., & McIlroy-Young, R. (2017). Introducing metaknowledge: Software for computational research in information science, network analysis, and science of science. *Journal of Informetrics*, *11*(1), 176–197. https://doi.org/https://doi.org/10.1016/j.joi.2016.12.005

Merton, R. K. (1965). *On the Shoulders of Giants; a Shandean Postscript.* New York: The Free Press.

Phinney, D. G. (2002). Building a consensus regarding the nature and origin of mesenchymal stem cells. *Journal of Cellular Biochemistry*, *85*(38), 7–12. https://doi.org/10.1002/jcb.10084

Robertson, J. A. (2010). Embryo Stem Cell Research: Ten Years of Controversy. *The Journal of Law, Medicine & Ethics*, *38*(2), 191–203. https://doi.org/10.1111/j.1748-720X.2010.00479.x

Robey, P. G. (1992). *Notice of Intramural Research Project*. Bethesda, Maryland.

Robey, P. G. (1993). *Notice of Intramural Research Project*. Bethesda, Maryland.

Robey, P. G. (1994). *Notice of Intramural Research Project*. Bethesda, Maryland.

Sacchetti, B., Funari, A., Remoli, C., Giannicola, G., Kogler, G., Liedtke, S., … Bianco, P. (2016). No Identical "Mesenchymal Stem Cells" at Different Times and Sites: Human Committed Progenitors of Distinct Origin and Differentiation Potential Are Incorporated as Adventitial Cells in Microvessels. *Stem Cell Reports*, *6*(6), 897–913. https://doi.org/10.1016/j.stemcr.2016.05.011

Sharpe, P. T. (2016). Dental mesenchymal stem cells. *Development*, *143*(13), 2273 LP-2280. https://doi.org/10.1242/dev.134189

Sugimoto, C. R., & Mostafa, J. (2018). A note of concern and context: On careful use of terminologies. *Journal of the Association for Information Science and Technology*, *69*(3), 347–348. https://doi.org/doi:10.1002/asi.24014

Tavassoli, M., & Friedenstein, A. (1983). Hemopoietic stromal microenvironment. *American Journal of Hematology*, *15*(2), 195–203. https://doi.org/10.1002/ajh.2830150211

Thor, A., Marx, W., Leydesdorff, L., & Bornmann, L. (2016). Introducing CitedReferencesExplorer (CRExplorer): A program for reference publication year spectroscopy with cited references standardization. *Journal of Informetrics*, *10*(2), 503–515. https://doi.org/https://doi.org/10.1016/j.joi.2016.02.005

Tuan, R. S., Boland, G., & Tuli, R. (2002). Adult mesenchymal stem cells and cell-based tissue engineering. *Arthritis Res Ther*, *5*(1), 32. https://doi.org/10.1186/ar614

Van Calster, B. (2012). It takes time: A remarkable example of delayed recognition. *Journal of the American Society for Information Science and Technology*, *63*(11), 2341–2344. https://doi.org/10.1023/B:SCIE.0000006881.30700.ea

van Eck, N. J., & Waltman, L. (2010). Software survey: VOSviewer, a computer program for bibliometric mapping. *Scientometrics*, *84*(2), 523–538. https://doi.org/10.1007/s11192-009-0146-3

Van Raan, A. F. J. (2004). Sleeping beauties in science. *Scientometrics*, *59*(3), 467–472. https://doi.org/10.1023/B:SCIE.0000018543.82441.f1

Vasiliev, J. M. (2008). No About I.M. Gelfand's Seminar. *Russian Journal of Developmental Biology*, *39*(6), 371–373. https://doi.org/10.1134/S106236040806012X

Wagner, W., & Ho, A. D. (2007). Mesenchymal Stem Cell Preparations—Comparing Apples and Oranges. *Stem Cell Reviews*, *3*(4), 239–248. https://doi.org/10.1007/s12015-007-9001-1

Wang, Y., Chen, X., Cao, W., & Shi, Y. (2014). Plasticity of mesenchymal stem cells in immunomodulation: pathological and therapeutic implications. *Nature Immunology*, *15*, 1009. Retrieved from http://dx.doi.org/10.1038/ni.3002

Winkler, W. E. (1990). String Comparator Metrics and Enhanced Decision Rules in the Fellegi-Sunter Model of Record Linkage., 354–359.

Wray, K. B., & Bornmann, L. (2015). Philosophy of science viewed through the lense of "Referenced Publication Years Spectroscopy" (RPYS). *Scientometrics*, *102*(3), 1987–1996. https://doi.org/10.1007/s11192-014-1465-6

Ziman, J. (2000). *Real science: What it is and what it means*. Cambridge University Press

When peer reviewers go rogue - Estimated prevalence of citation manipulation by reviewers based on the citation patterns of 69,000 reviewers

Jeroen Baas[1] and Catriona Fennell[2]

[1] *j.baas@elsevier.com*
Elsevier B.V., Radarweg 29, 1043 NX Amsterdam (The Netherlands)

[2] *c.fennell@elsevier.com*
Elsevier B.V., Radarweg 29, 1043 NX Amsterdam (The Netherlands)

Abstract

There is anecdotal information available that some reviewers attempt to increase their citation counts by using the peer review process, adding references to reviewed publications. There have been studies around citation coercion from the perspective of journal editing and boosting of journal indicators. This study builds further on that work, with a different angle: by measuring excessive citation manipulation at the level of reviewers. In order to assess the extent of this behaviour, access to a large pool of peer-review records is required: connections between authors and reviewers, connections between reviewers and reviewed work and so forth. This study explores this area in two phases. In phase one we detect the overall patterns of citations from reviewed material to reviewers, and assign a value to the proportion of citations originating from reviewed work. The second phase further explores the citation patterns, by taking the outliers from phase one and identifying the citations that have been added during the review process. We find in the results that highly suspicious cases of this behaviour can be successfully detected and that the scale of suspicion of clear misconduct behaviour is relatively limited (0.79%).

Introduction

Researchers are increasingly pressured to publish ("publish or perish"), and more recently to maximise citations to their published work (Plume & Van Weijen, 2014) (Anderson, 2007). Journal-level metrics are still used today in researcher evaluation, even though reports such as the metric tide report (Wilsdon, et al., 2015) identified risks and misrepresentations using such metrics for individual evaluation. Since its launch in 2012, more than 700 organisations have committed to the Declaration on Research Assessment (DORA) which recommends that evaluation of scholarly research outputs be improved and rely less on journal-level metrics (DORA, 2012).

This has led to an increased focus on other metrics with a more individual measurement component: article-citations. Examples of such metrics are the h-index, aggregated field normalized citation indicators and percentiles. These metrics are not without debate either: self-citations as a means to game citation metrics is a returning topic at the presentation of citation-based evaluation studies (Shema, 2012). Researchers can have multiple incentives to increase their citations; it builds esteem or "vanity" (Franck, 1999) and is used as supportive information in many assessment exercises (REF 2021, 2017) (Preti, 2017) (Abramo & D'angelo, 2015) (ERA 2018 Submission Guidelines, 2017). Next to these extrinsic reasons, researchers have many intrinsic incentives to increase citations, such as dissemination of their work and the use of available resources. A pursuit of citations cannot inherently be classified as ethical or unethical.

The use of journal-level indicators for research assessment may lead to incentives for journals to increase their scores: authors seek out higher scored journals to publish in and journals seek submissions. This has led to numerous studies into citation stacking and citation cartels (Wilhite & Fong, 2012) (Hugget, 2013) (Fong & Wilhite, 2017) (Davis, 2017). These studies reviewed the coercion introduced by editors to add citations to submitted publications in order to increase

the citation score of a journal. This may be the same journal as where the work is submitted, or in some cases, more advanced cartels where journals cite each other in order to stay under self-citation thresholds introduced as a counter measure by journal-metric providers. There are several case studies and small scale studies available that deal with reviewer induced citation coercion to work of the reviewer (Resnik, Gutierrez-Ford, & Peddada, 2008) (Ioannidis, 2015) (Thombs, et al., 2015)

Detecting citation manipulation at the level of individual reviewers is more complicated to achieve. The first challenge is the conceptual difficulty of defining the intent of the reviewer. Reviewers are often for good reasons experts in the area of the reviewed material and as such, often have relevant citable work available. For the scope of this study, the extent to which these patterns can be labelled as excessive are evaluated. The second challenge is the technical access to linked data. In addition to access to the reviewed material, information on which reviewers contributed to the reviews of a particular article is required. Information about the references in reviewed work is also required: was the reviewer cited in the manuscript? This requires disambiguation at the level of authors, while citation manipulation at journal levels can be detected using only journals. Furthermore, in order to contextualize these numbers derived from the connections, information about the universe of publications and citations of a reviewer is required. For this a connection between the reviewer and a publication database' author profile is needed.
Using a dataset consisting of 69,000 reviewers, their reviews for Elsevier-published journals during a 3-year period (2015-2017) and the Scopus database with its over 12 Million author profiles, this study aims to explore and identify the prevalence of citations from reviewed work to reviewers. Furthermore, for the high-end outliers of reviewers with a high degree of reviewer-citations, it investigates the degree to which these citations have been added during the publication process.

Background
Reviewers are experts in an academic field who critically appraise the accuracy, quality and robustness of scientific articles written by their peers in order to improve the clarity and quality of the article (Ware, M. 2013). Reviewers typically do so at the request of a journal editor to help determine whether a submitted manuscript is suitable for publication in that journal, known as "pre-publication peer review". In the vast majority of peer-reviewed journals, reviewers' identities are not shared with authors either before or after publication.
In this study, all reviewers used in analysis have also acted as authors of papers included in Scopus.

Authors
In this research, *authors* refers to authors of the manuscript that was submitted for review. Authorship is also relevant for other publications from the same group of reviewed-authors, in order to assess self-citations: the citation links between publications that share one or more of the same authors. Authors can be reviewers on other articles, but do not necessarily have to be reviewers.

Reviewer citations
When a citation is pointing to a publication that is authored by one of the reviewers of the reviewed paper, this is referred to as a *reviewer citation*. These citations may be the consequence of an intervention of the reviewer in the review process, but it is also possible – or even probable- that the reviewer citation was already present in the original manuscript. Some editors routinely identify potential reviewers for submissions from the authors of papers

cited in the submission. In those cases, the reviewer citation is not added during review. Differentiating between reviewer citations that were present in the original manuscript and those added during the publication process can be complex. Original manuscripts have not yet undergone the quality checking, structuring or tagging that takes place before final publication. Complete data on all versions of manuscripts may not be retained indefinitely.

Methods

Reviewer data is collected from two submission systems used by Elsevier-published journals: EES and EVISE. Due to the background of EES, the system in itself is comprised of many databases that each aggregate to a journal. A result of this architecture is that reviewers will have multiple records if they have reviewed work submitted to EES and EVISE, and also if they have reviewed work submitted to different journals under EES alone. The extracted dataset consists of 1,011,826 manuscripts published between 2015 and 2017, for 985,405 of which the original references are still available. For each of these publications, reviewer metadata records associated with the manuscript are identified, with a unique total of 1,264,633 records. Data available in the submission system are ORCID, email addresses and pre-identified author IDs.
Scopus data is used to analyse the citation patterns of authors. Scopus is an abstract and citation database provided by Elsevier (Schotten, el Aisati, Meester, Steiginga, & Ross, 2017) (Berkvens, 2012). An indexed version of the Scopus database with an extraction date of 02-November-2017 is used, based on the raw full XML dataset that is similarly indexed by the Scopus.com platform for search and retrieval purposes.
There are three important matching procures used in this experiment. First, publications that have been reviewed and identified as the dataset used for evaluation need to be matched to Scopus records. This is achieved using a two-step link for EES. An identifier for the record within the scope of the EES database, accompanied by the database ID is linked to a PII, which is the ID as used by the ScienceDirect platform on which the accepted paper is published. Similarly for EVISE, a direct mapping exists between a record-ID and a PII. Both resulting PII mappings are then mapped to Scopus records using the PII field as captured by the raw Scopus XML data.

Second, it is important to map the reviewers to authored publications. Authored publications are the citing target that will be tracked. Mapping is achieved by first matching reviewers to author profiles in Scopus, and subsequently retrieving the publications as captured by that author profile. Author profiles in Scopus are a combination of curated and system generated profiles. All author profiles are originally generated by an “author profiling” algorithm (Schotten, el Aisati, Meester, Steiginga, & Ross, 2017). This results in profiles that are optimized for precision (publications merged in a profile belong to one and the same person) over recall (all publications of the same person are merged into one profile). Precision is appropriate for this experiment as well: precision errors would lead to potential underestimations of reviewer citations. Recall errors are less severe: lack of identification of the authored work of the reviewer means that the experiment will not pick up on all reviewer citations. it means measuring the experiment across a subsample of the work of the reviewer: not identifying . Nevertheless, high rates of precision and recall are desirable for a robust statistic. The current reported precision and recall by Scopus is 98%, at an average recall of 93.5%.

Many profiles in Scopus have undergone manual curation. There are several processes that could lead to human intervention of author profiles. The most well-established process is through ORCID, an open, non-proprietary registry of unique, persistent author identification codes. ORCID offers researchers facilities to claim and register publications, funding and other

works (What is ORCID?, 2018). As part of that process, a flow through Scopus is offered resulting in a curated profile on Scopus. Similarly, researchers have the ability to use the author feedback wizard (APW). Another process resulting in curated profiles is a commercial service offered by Elsevier as part their Pure offering, called Profile Refinement Service. All these efforts combined lead to a current count of estimated 1,5 million author profiles that have undergone manual curation. The manual curation count is determined by a flag in the XML data of the produced Scopus Author Profile records, and the number above is derived by counting the records with this flag. Reviewers may be linked to multiple author profiles, as there is the possibility that different author records for the same person have been created: the system generated profiles favour precision over recall, which in uncertain cases results in more than one profile for the same person. Similarly, an author may have multiple reviewer records: referees that have reviewed for multiple journals using EES as well as both on EVISE and EES will have multiple reviewer records in the database. The result of this cardinality in the data model is a many-to-many relationship between reviewer records and author profiles. This is important for the experiment, as the experiment benefits from a maximum recall of both the cited (authored) and citing (reviewed) works. The profiles in Scopus and reviewer records are mapped using ORCID, and email addresses. Email addresses are only directly mapped between authors and referees if the email address is not used by more than 3 author profiles. In some cases, email addresses link back to a lab address rather than a personal email address. This results often in assignment of the same email address to multiple different authors profiles, and therefore it is desirable to not include those cases. Mapping reviewer records to author profiles is possible for 875,692 records, and after grouping of records based on shared profiles, the population of grouped reviewer records is 506,614.

The third mapping required for the experiment, is between references on the first version of the submitted manuscript and Scopus. Citations from the final published version (version of record) to Scopus articles are inherent to the Scopus database and directly linked using internal IDs (Scopus EIDs). However, the first version of a manuscript does not contain cleaned and linked references, but only contains raw text (see Figure 1). In order to assess which of the references were not present in the first version, and thus added during review, it is needed to identify which of the references existed in the first version. Other references in the final version are assumed to be added during review. Linking raw, unstructured citations to a database can be achieved with different approaches. Libraries and tools exist that allow extraction of structure from unstructured references (Chenet, 2017). In this experiment, structure of the reference itself is not relevant: a link between a reference and the best potential record in Scopus is most important. Precision of the match is less important for cited references that are not covered by Scopus; this stage is designed to identify added references, and a false positive match to Scopus on a reference in the first version is quite unlikely a match to a reference that was added to the final version. Recall is important at this stage, as each original reference not identified in Scopus will appear as a reference added during review. To build the link between the raw text of a single reference and Scopus, a search strategy with fuzzy logic is used. For that purpose, all fields from Scopus that are usually found in references (such as author names, title, year of publication, journal title, volume, issue, pages, DOI etc) or combined into a single searchable indexed field in a SOLR index (Apache Solr Reference Guide, 2015). Each to-be-matched reference in turn is converted into an OR search string and results are sorted by relevance. Next, the top 5 results are evaluated in a post-process to assess validity of the result. This is achieved by heuristic scoring of different elements using a point-based system. For example, presence of the Scopus record candidate publication year in the source string results in a .15 score. More points are accumulated when other fields match. If over 90% of authors for a paper with more than 5 authors are linked, an additional bonus point is added, and something similar happens

for matching of more words of a longer article title. A random sample of 114 unstructured references was manually inspected to test precision and recall of the linking methodology. Precision of matched records was 0.97, recall 0.97, leading to an F-score of 0.97.

```
1. APHA (American Public Health Association), 2005.
Standard Methods for the Examination of Water and
Wastewater, 21st ed. Byrd Prepress, Washington.
2. Arar, E.J., Collins, G.B., 1997. "In vitro"
determination of clorophyll a and pheophytin a in marine
and freshwater algae by fluorescence (EPA method 445.0).
Environmental Protection Agency, Cincinnati
```

Figure 1 Example of unstructured references

In this analysis, the Scopus database is used, consisting of over 70 million articles. The reviewer material is limited to submissions to the Elsevier EES and EVISE submission systems. This means that citations from other publishers that can be detected in the Scopus database cannot be classified definitively for other publishers than Elsevier. For this reason, the citation indicators used to benchmark proportions of publications and citations per author are limited to Elsevier published material only, so that a reviewer's citation proportion is not influenced by the difference in publisher distributions across authors.

The experiment is set up in two phases. In the first phase, the aim is to identify potential outlier reviewers with a high degree of citations originating from reviewed material. In order to achieve this, the first two mappings (reviewed work to Scopus and reviewers to Scopus authored papers) are applied to the EES, EVISE and Scopus datasets. From this linked dataset, statistics are extracted per reviewer as outlined in Table 1.

Table 1 Variables measured in phase one

Variable name	*Description*
count reviewed documents in Scopus	Number of documents reviewed, that are mapped to an item in Scopus. This omits reviewed material not indexed in Scopus.
sum of references to reviewer	Number of references in reviewed documents, citing the reviewer
count reviewed documents in Scopus citing reviewer	Number of reviewed document that cite the reviewer
REV authored citsElsPeriod	Number of citations to documents authored by reviewer originating from publications during the period with Elsevier
REV authored selfCitsElsPeriod	Number of self-citations to documents authored by reviewer originating from publications during the period with Elsevier

These statistics are further combined into proportion indicators: (1) the proportion of citing documents, as expressed by: [count reviewed documents in Scopus citing reviewer] / [count reviewed documents in Scopus]. (2) The proportion of the citation count of the reviewer, that is originating from reviewed articles: [sum of references to reviewer] / [REV authored citsElsPeriod]. This proportion is restricted in the denominator (citation count of reviewer) to only citations occurring during the evaluation period of the study, and originating from other Elsevier' published articles. Restricting the scope is required as the numerator is within the same scope and therefore only this scope can be explained in the study. (3) The proportion of the citation count of the reviewer, where none of the citations are self-citations, that is

originating from reviewed articles: [sum of references to the reviewer] / ([REV authored citsElsPeriod] - [REV authored selfCitsElsPeriod]). The exploration in this study deals with the prevelance of citation manipulation, and therefore the proportion excluding self-citations is a strong indicator of the proportion across citations that are not already under direct control of the reviewer. The excluded self-citations for this indicator are in turn not considering self-citations that are also reviewer citations. This can be the case when for instance a reviewed article cites the reviewer, but the reviewer and the reviewee have co-authored a publication that is cited. This in itself is a potential issue that is not considered in this study and it should be noted that the reviewer selection systems are already equipped with co-author detection to minimize this risk of conflict of interest.
Outliers on the proportional indicators for citation counts (2 and 3) are classified by use of the upper inner fence; 1.5 inter quartile range from the median in a normal distribution. The citation count proportions are log-normal distributed. The proportional indicator for citing reviewed documents (1) is not normal distributed, as assignment of reviewers is often based on reference lists: the indicator may partially be explained by assignment and not action of the reviewer. In addition, due to the distribution of number of reviewed publications (expected mode of 5) the proportion has few discretionary values at high frequency (mode of 5 means 0.2, 0.4, 0.6, 0.8 and 1.0). Classification of outliers for this indicator is therefore based on heuristics after inspection of the histogram.

In order for this phase to be based on a robust set of data, with sufficient opportunity for authored and reviewed works by reviewers to be linked, only reviewers with 5 or more authored publications in Scopus and having reviewed 5 or more articles are considered. This reduces the population of 506,614 reviewers in total, to 69,096. Out of these, 14,275 (20%) have no citations from reviewed articles to authored articles and are excluded from the analysis. This further reduces the population of reviewers analysed to 54,821.
The second phase uses the subset of outliers identified in phase 1. From the reviewed articles containing one or more citation to the reviewer, the references from the original manuscripts are retrieved and linked using the third mapping process outlined above. For some original manuscripts, the reference data is no longer available. Only reviewers for whom 50% or more of the reviewed articles with available reference data are included in phase 2. This is to prevent a too low sample of references across reviewed work from clouding the results. Phase two measures one variable: the proportion of citations to the reviewer, that have been added during the review process. A threshold of 50% is set as suspicious. This means that reviewers are considered suspicious if 50% or more of the citations to the reviewer are considered to be added during review.

Results
Input to phase 1 are 54,821 reviewers: those meeting all criteria of having 5 or more authored publications, reviewed 5 or more publications of which at least one cites the reviewer. The frequency distribution of the proportion of reviewed documents citing the reviewer show a non-normal distribution. From the histogram, we draw the conclusion that an outlier section exists on the far end of the scale: 100% of reviewed documents citing the reviewer. This represents 1,612 reviewers.

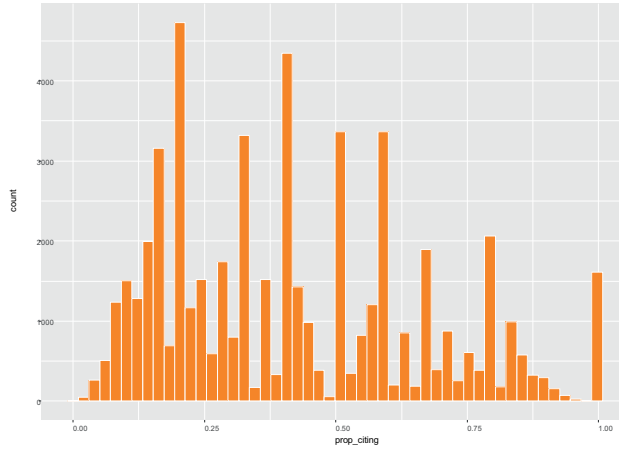


Figure 2 histogram of count of reviewers by proportion of reviewed documents citing the reviewer(proportion indicator 1)

The second indicator measures the proportion of citations explained by reviewer citations. This is a log-normal distribution (see Figure 3). Q1 is at -5.43 (0.43%), Q2 (median) at -4.52 (1.08%) and Q3 at -3.64 (2.62%). The inter quartile range is 1.79. The upper inner fence is at -0.95 (38.4%), resulting in a selection of 108 reviewers.

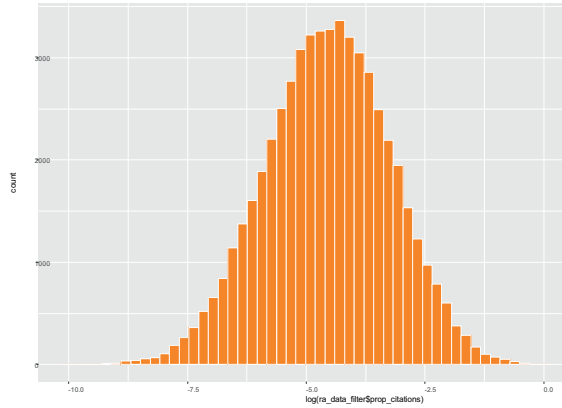


Figure 3 Reviewers by proportion of citations originating from reviewed work (proportion indicator 2)

The third indicator measures the proportion of non-self-citations explained by reviewer citations. This also is a log-normal distribution (see Figure 4). Q1 is at -5.2 (0.54%), Q2 (median) at -4.26 (1.4%) and Q3 at -3.33 (3.5%). The inter quartile range is 1.87. The upper inner fence is at -0.52 (59.3%), resulting in a selection of 69 reviewers.

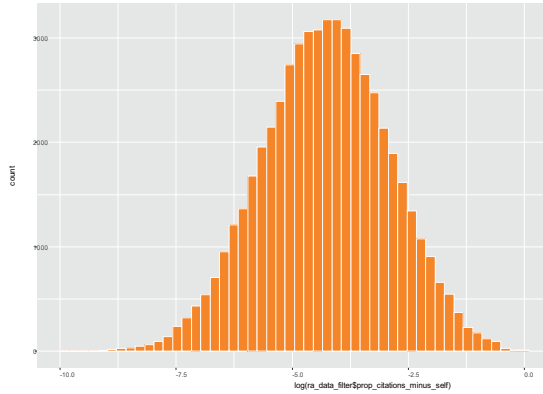


Figure 4 Reviewers by proportion of non-self-citations explained by reviewer citations (proportion indicator 3)

Combining the outlier selections from the three metrics results in a set of 1,734 reviewers for analysis in phase 2. Out of these, for 1,041 reviewers, 50% or more of the reviewed original manuscripts contain one or more references (see Figure 5). Only those reviewers are further analysed in this phase, and only the publications with an original reference list were analysed.

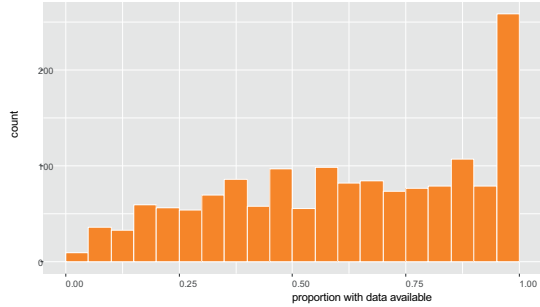


Figure 5 Reviewers by proportion of reviewed manuscripts with original reference list

Out of these reviewers, in 260 cases 50% or more of the reviewer citations in the reviewed article were not found in the original manuscript (see Figure 6).

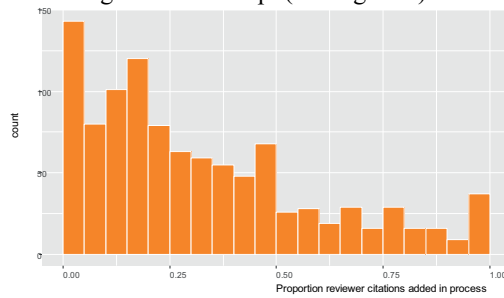


Figure 6 Reviewers by proportion of the reviewer citations in the reviewed article not found in the original manuscript

Extrapolating the 260 (25% of 1,041) reviewer cases to the 1,734 reviewers leads to a count of 433, a proportion of 0.79% of 54,821 reviewers.

Discussion

Comparable analyses on reviewer citation manipulation data at this scale are not available. This study aims to provide the necessary benchmarks and early prevalence indications of citation manipulation originating from review. The identification of 433 reviewers with high suspicion (0.79%) indicates a low prevalence of systematic citation manipulation. Absence of similar datasets means that findings in this study cannot be qualified as high or low in comparison with other sources. An important limitation of this study is that the underlying intent of suggested citations by reviewers cannot be measured. The statistical analysis highlights suspicious profiles, that need further human review in order to also assess the relevance of the added citations and the likely intent of the reviewer.

Detailed human assessment and investigation of the most suspicious cases highlighted initial patterns which may help to explain why this behaviour was not detected and curtailed before publication of the articles. Citation recommendations within the reviews were often presented in specific styles and formats which reduced the likelihood of editors easily detecting them as erroneous recommendations. For example: references were often mentioned in reviewer reports with the author names of the references omitted, obscuring the fact that all papers being recommended were authored by the reviewer. There is potential for development of functionality within editorial systems to help alert editors to suspicious numbers or types of citation recommendations. From the author perspective, since authors are typically blinded to reviewers' identities, while they may suspect that the citations they are being asked to add relate to the reviewer's own papers, the authors cannot know definitively. Even when authors find the citation recommendations to be erroneous or irrelevant to their paper, they may decide that adding them is a relatively painless way to respond to the reviewers' comments. Given that reviewers may ask for far more challenging revisions to the submission that authors prefer to decline (for example requiring additional experiments to be conducted), authors may see the addition of questionable citations as a minor concession. In one case where a reviewer had submitted 120 reviewer reports including requests to add multiple irrelevant citations in each report, only four authors refused to add the citations during revision of their paper.

An early exploration of a large dataset also means that not all angles and approaches have been considered. We suspect that due to different citation pattern behaviour between fields (Schubert, Glänzel, & Braun, 1983), patterns between fields of this same prevalence may differ. It would be interesting to explore both the prevalence between fields (i.e. based on the outlier detection across the entire corpus) as well as within fields to see if there are different observations across fields in the relation between first outlier detection (phase 1) and the proportion of reviewer citations added in the process (phase 2). One challenge with such an approach would be to properly characterize reviews as belonging to a certain field; the most suspicious reviewers operate across disciplines in many journals and assigning these to a field may prove challenging. Another element for further study is the robustness of the chosen parameters and threshold in phase 1 and phase 2, and the stability of the results.

In a recent study, the effects of publishing review reports have been studied (Bravo, Grimaldo, López-Iñesta, Mehmani, & Squazzoni, 2019). Only 8.1% of the referees agreed to reveal their identity in the published report, which prevents systematic identification of citation

manipulation. Public reviews could still partially affect these issues: reviewers wrote more constructive reports in the case of public reviews. Public reviews may additionally be relevant if research will suggest they are being viewed beyond the authors and editors, another related area of future research. It is likely that the eventual discovery of large-scale reviewer citation manipulation in several soil science journals was partially enabled by Copernicus Publications' practice of publishing reviews (van Groenigen, et al., 2018).

Conclusions

This exploratory study suggests that prevalence of citation manipulation by means of review is low for the corpus analysed (0.79%). Identification of reviewing profiles leads to a suspicion only, as the underlying intent of reviewers cannot be proven mathematically. For example, all suggested citations may have been fully relevant to the article. A further review of all profiles is needed to determine if the behaviour can be classified as unethical.

Identifying unethical behaviour in peer review also brings complicated decisions to be made: what should and can journals do if editors judge that misconduct has taken place? In the comparable situation of an author committing misconduct, the journal ultimately has the option to correct the literature by retracting the paper, which is a public and permanent warning to readers that the paper is no longer reliable (COPE, 2009). A record of the retraction is also typically fed through to secondary databases such as PubMed, Web of Science and Scopus. However, the publishing industry currently has no comparable practice or technical mechanism for retracting an unreliable citation. While journals could publish corrections to the articles informing readers that certain references within the article are erroneous, those corrections would have no impact on the citation records within secondary databases at the author, journal or institutional level.

Without such a mechanism available as a deterrent, journals can discourage such inappropriate behaviour by other means: making it explicit to reviewers at the time they agree to review that this practice is unacceptable (Elsevier Ethical guidelines, 2019), by no longer entrusting repeat offenders with future submissions to review and by sharing evidence with their institutes, as appropriate. Institutes can play a powerful role in censuring such behaviour. Unfortunately, in some cases the institute may not act or the reviewer may be independent of any academic or commercial institute. While journals may keep records to avoid sending submissions to such reviewers, due to privacy and legal concerns, such records are unlikely to be shared between different journals and publishers. In all cases of sharing of personal details around unethical behaviour, complex questions need to be answered; how strong is the evidence of intent? Is there a clear rule that the offender should have known about? If actions are taken, is there a place where the researcher can appeal? Does the researcher remain excluded indefinitely or can the researcher make amends?

Recently, a case-study investigating a specific reviewer citation coercion case in a journal identified the need for similar guidelines for reviewers, authors, journals and editors (Wren, Valencia, & Kelso, 2019). However, the question remains as to how journals should act beyond best attempts at prevention and detection of reviewer citation coercion patterns. The current lack of serious and permanent consequences limits journals' ability to deter such behaviour and to transparently correct the citation record when it has been detected.

In terms of the broader research community, a code of conduct applied by local scientific systems may have sections dealing with specifics around review. An example is The Dutch universities code of conduct 2018, section 3.5.48: "*Do not use the system of peer review to generate additional citations for no apparent reason, with the aim of increasing your own or other people's citation scores ('citation pushing')*." (KNAW; NFU; NWO; TO2-federatie; Vereniging Hogescholen; VSNU, 2018).

The identification of potential unethical behaviour in peer review shown in this paper highlights new challenges and unresolved questions for publishers in terms of how to effectively deter and detect such misconduct, and how to responsibly correct the literature upon identification.

CRediT statement

Jeroen Baas: Conceptualization; Formal analysis; Methodology; Writing – original draft; Writing – review & editing

Catriona Fennell: Conceptualization; Project administration; Writing – review & editing

Acknowledgments

The authors would like to acknowledge and thank Jan Willem van Groenigen, Theo Jetten, Ellen Fest and Coen Ritsema, Wageningen University for their guidance; and Elizabeth Ash, Freek Bedaux and Ramsundhar Baskaravelu from Elsevier for assistance in acquiring the required reviewer datasets.

References

Abramo, G., & D'angelo, C. (2015, December 1). An assessment of the first "scientific habilitation" for university appointments in Italy. *Economia Politica, 32*(3), 329-357. doi:10.1007/s40888-015-0016-9

Anderson, M. e. (2007). The perverse effects of competition on scientists' work and relationships. *Science and Engineering Ethics* , 13: 437–461.

Apache Solr Reference Guide. (2015, February 19). Retrieved from Archives of the Apache Software Foundation: http://archive.apache.org/dist/lucene/solr/ref-guide/apache-solr-ref-guide-5.0.pdf

Berkvens, P. (2012). *Scopus Custom Data Documentation.* Retrieved from http://ebrp.elsevier.com/pdf/Scopus_Custom_Data_Documentation_v4.pdf

Bravo, G., Grimaldo, F., López-Iñesta, E., Mehmani, B., & Squazzoni, F. (2019). The effect of publishing peer review reports on referee behavior in five scholarly journals. *Nature Communications, 10*, 322.

Chenet, M. (2017). *Identify and extract entities from bibliography references in a free text.* Master Thesis, Universiteit Twente, Human Media Interaction. Retrieved from https://essay.utwente.nl/73817/1/chenet_MA_EEMCS.pdf

COPE. (2009). Retraction Guidelines. *Committee on Publication Ethics*.

Davis, P. (2017, May 13). *How Much Citation Manipulation Is Acceptable?* Retrieved November 10, 2018, from Scholarly Kitchen: https://scholarlykitchen.sspnet.org/2017/05/30/how-much-citation-manipulation-is-acceptable/

DORA. (2012). Retrieved December 19, 2019, from San Francisco Declaration on Research Assessment: http://www.ascb.org/files/SFDeclarationFINAL.pdf

Elsevier Ethical guidelines. (2019). *Duties of Reviewers Guidelines*. Retrieved May 15, 2019, from https://www.elsevier.com/about/policies/publishing-ethics

ERA 2018 Submission Guidelines. (2017). Retrieved November 13, 2018, from Australian Research Council: https://www.arc.gov.au/excellence-research-australia/key-documents

Fong, E. A., & Wilhite, A. W. (2017). Authorship and citation manipulation in academic research. *Plos ONE, 12*(12), e0187394. doi:10.1371/journal.pone.0187394

Franck, G. (1999, October 1). Scientific Communication--A Vanity Fair? *Science, 286*(5437), 53-55. doi:10.1126/science.286.5437.53

Hugget, S. (2013). Journal bibliometrics indicators and citation ethics: a discussion of current issues. *Atherosclerosis, 230*(2), 275-277.

Initial decisions on the Research Excellence Framework 2021 (REF 2017/01). (2017, September 1). Retrieved November 13, 2018, from REF2021: https://www.ref.ac.uk/media/1050/ref2017_01.pdf

Ioannidis, J. P. (2015). generalized view of self-citation: direct, co-author, collaborative, and coercive induced self-citation. *Journal of Psychosomatic Research, 78*(1), 7-11. doi:10.1016/j.jpsychores.2014.11.008

KNAW; NFU; NWO; TO2-federatie; Vereniging Hogescholen; VSNU. (2018). *Nederlandse gedragscode wetenschappelijke integriteit.* Retrieved from VSNU: https://doi.org/10.17026/dans-2cj-nvwu

Plume, A., & Van Weijen, D. (2014). Publish or perish? the rise of fractional autor. *Research Trends, 38*.

Preti, A. (2017, February 25). Italian Abilitazione Scientifica Nazionale. *The Lancet, 389*(10071), 803-804. doi:10.1016/S0140-6736(17)30513-5

Resnik, D. B., Gutierrez-Ford, C., & Peddada, S. (2008). Perceptions of Ethical Problems with Scientific Journal Peer Review: An Exploratory Study. *Science and Engineering Ethics, 14*(3), 305-310. doi:10.1007/s11948-008-9059-4

Schotten, M., el Aisati, M., Meester, W., Steiginga, S., & Ross, C. (2017). A Brief History of Scopus: The World's Largest Abstract and Citation Database of Scientific Literature. In F. Cantu-Ortiz, *Research Analytics. Boosting University Productivity and Competitiveness through Scientometrics.*

Schubert, A., Glänzel, W., & Braun, T. (1983). Relative citation rate: A new indicator for measuring the impact of publications. *First national conference with international participation on scientometrics and linguistics of scientific text.* Varna.

Shema, H. (2012, July 24). *On Self-Citation*. Retrieved November 12, 2018, from Scientific American: https://blogs.scientificamerican.com/information-culture/on-self-citation/

Thombs, B. D., Levis, A. W., Razykov, I., Syamchandra, A., Leentjens, A. F., Levenson, J. L., & Lumley, M. A. (2015). Potentially coercive self-citation by peer reviewers: A cross-sectional study. *Journal of Psychosomatic Research, 78*(1), 1-6. doi:10.1016/j.jpsychores.2014.09.015

van Groenigen, J. W., Agnelli, A., Bai, J., Capowiez, Y., Cayuela, M., Kögel-Knabner, I., . . . Vepraskas, M. (2018). Citation stacking in soil science articles: our response to the open letter by concerned early-career soil scientists. *Geoderma, 328*, 119-120. doi:10.1016/j.geoderma.2018.03.023

What is ORCID? (2018). Retrieved November 13, 2018, from ORCID: https://support.orcid.org/hc/en-us/articles/360006897674

Wilhite, A., & Fong, E. (2012). Coercive citation in academic publishing. *Science, 335*(6068), 542-543. doi:10.1126/science.1212540

Wilsdon, J., Allen, L., Belfiore, E., Campbell, P., Curry, S., Hill, S., . . . Johnson, B. (2015). *The Metric Tide: Report of the Independent Review of the Role of Metrics in Research Assessment and Management.* HEFCE. doi:10.13140/RG.2.1.4929.1363

Wren, J., Valencia, A., & Kelso, J. (2019, January 30). Reviewer-coerced citation: Case report, update on journal policy, and suggestions for future prevention. *Bioinformatics*, btz071. doi:10.1093/bioinformatics/btz071

Bibliometrics for collaboration works

Paolo Rossi[1], Alessandro Strumia[2], and Riccardo Torre[3]

[1] *paolo.rossi@unipi.it*
Dipartimento di Fisica dell'Università di Pisa, Largo Bruno Pontecorvo 3, 56127 Pisa (Italy)

[1] *alessandro.strumia@unipi.it*
Dipartimento di Fisica dell'Università di Pisa, Largo Bruno Pontecorvo 3, 56127 Pisa (Italy)

[3] *riccardo.torre@cern.ch*
CERN, Theory Division, Esplanade des Particules 1, 1211 Genève (Switzerland)
INFN, Sezione di Genova, Via Dodecaneso 33, 16146 Genova (Italy)

Abstract

An important issue in bibliometrics is the weighing of co-authorship in the production of scientific collaborations, which are becoming the standard modality of research activity in many disciplines. The problem is especially relevant in the field of high-energy physics, where collaborations reach 3000 authors, but it can no longer be ignored also in other domains, like medicine or biology. We present theoretical and numerical arguments in favour of weighing the individual contributions as $1/N_{\mathrm{aut}}^{\alpha}$ where N_{aut} is the number of co-authors. When counting citations we suggest the exponent $\alpha \approx 1$, that corresponds to fractional counting. When counting the number of papers we suggest $\alpha \approx 1/3 - 1/2$, with the former (latter) value more appropriate for larger (smaller) collaborations.

Introduction

In many research fields, scientific collaboration has become the standard way of operating, and moreover, due to the increasing complexity of the problems to be faced, the number of scientists with different competences involved in each single collaboration is increasing. In the extreme case of high energy physics numbers have already reached the four-digit level, but in many other domains, like medicine of biology, it is not unusual to find two-digit collaborations.

In the context of bibliometrics this hyper-authorship phenomenon pose a very important question, concerning the individual degree of property that must be assigned to the authors of a common scientific article and to the citations received by a paper. On one side, it is rather clear that attributing the full credit of a paper to each of the author is mystifying and (if adopted by policymakers) tends to encourage fictitious collaborations, because of the obvious competitive advantage resulting from the much larger number of articles that a collaboration may produce in the same amount of time in comparison with an isolated author. Moreover, also the number of citations received is strongly correlated with the typical dimensions of the collaborations operating in a given field of research.

The choice of fractionally counting papers by attributing a $1/N_{\mathrm{aut}}$ weight to each of the N_{aut} co-authors of a paper would imply a strong penalty and (if adopted by policymakers) would discourage collaborations, due to the obvious impossibility for N_{aut} authors to produce N_{pap} papers in the same time in which a single author may produce a single paper.

A reasonable compromise between these two extreme (and conceptually wrong) choices would therefore be quite welcome.

The bibliometric literature documents that collaboration papers tend to have higher impact than single-authored papers. Beaver (1986) studied physics, finding that co-authored research tends to be of higher quality than solo research. Bordons, Garcia-Jover, and Barrigon (1993) studied Spanish publications in pharmacology and pharmacy finding that internationally co-authored documents have higher impact than the remaining collaborative documents or non-collaborative ones. Avkiran (2013) found that collaboration leads to articles of higher impact

in finance, up to 4 collaborators. Gazni and Didegah (2011) found a significant positive correlation between the number of authors and the number of citations in Harvard publications. Hsu and Huang (2011) considered 90k articles in natural sciences, finding that the average number of citations scales as $N_{\mathrm{aut}}^{1/3}$ (data extend up to about 10 co-authors), up to wide fluctuations. Lee and Bozeman (2005) found that the number of peer-reviewed journal papers is strongly and significantly associated with the number of collaborators, unlike the number of fractionally-counted papers. Katz and Hicks (1997) studied how the average number of citations per paper varies with different types of collaborations. See also the works of van Raan (1997), Sooryamoorthy (2009) and Birnholtz (2006).
We will better quantify and understand the bibliometric output of collaborations.
General theoretical arguments concerning scale-free systems suggest that the scientific productivity of collaborations and the corresponding frequency distribution of citations should show some, at least approximate, power law dependence on N_{aut}. Empirical evidence appears to support these arguments. Finding the most appropriate exponents for these scaling laws would offer the possibility of weighting the production of collaborations in the bibliometric estimate of the (quantitative) value of their results in such a way as to discourage adaptive and opportunistic behaviours while encouraging more appropriate practices in the indication of co-authorship.
In Section 2 we develop and present some theoretical arguments in favour of weighing the individual contributions to a single paper as $1/N_{\mathrm{aut}}^{\alpha}$, where $\alpha \approx 1/3 - 1/2$, with the former (latter) value more appropriate for larger (smaller) collaborations. When counting overall citations we suggest the exponent $\alpha = 1$, corresponding to fractional counting. Fractional counting has been extensively discussed in the literature. For instance, it has been considered in the context of metrics and rankings by Aksnes, Schneider, and Gunnarsson (2012), Bouyssou and Marchant (2016), Carbone (2011), Egghe (2008), Hooydonk (1997), Leydesdorff and Bornmann (2010, 2011), Leydesdorff and Opthof (2010), Leydesdorff and Shin (2011), Rousseau (2014), Strumia and Torre (2019), and in the context of constructing research networks by Leydesdorff and Park (2017), Perianes-Rodríguez, Waltman, and van Eck (2016). Fractional counting gives an intensive quantity: this means, for example, that the total index of the European Union is the sum of its members, unlike what happens if full counting is adopted (see e.g. the discussions by Gauffriau (2017), Gauffriau, Larsen, Maye, Roulin-Perriard, and von Ins (2008), Strumia and Torre (2019), and Waltman and van Eck (2015)).

In Section 3 we analyze empirical data concerning a very large number of collaborations active in fundamental physics, where the range of available values of N_{aut} allows for sufficiently convincing estimates of the exponents describing the dependence on N_{aut} of the total number of papers and of the mean and total number of citations.

A theoretical approach

Scaling

The behaviour of collaborations with N_{aut} authors can be viewed as scale-free phenomena for a wide range of values of N_{aut}. Any upper limit on N_{aut} would be sufficiently large to exclude any sensible effect on the equilibrium distributions in the range of values we are interested to explore (3 to 4 orders of magnitude).
We therefore expect that the various indices N_I that characterise bibliometric outputs of collaborations are distributed at equilibrium following a power-law behaviour, which we parametrise as follows

$$\langle N_I \rangle = C_I N_{\mathrm{aut}}^{p_I}, \tag{1}$$

where C_I and the powers p_I are constants. For example this applies to the number of papers $N_{\mathrm{pap}} = C_{\mathrm{pap}} N_{\mathrm{aut}}^{p_{\mathrm{pap}}}$ of papers produced (in a definite amount of time) by a scientific collaboration and to the average number of citations per paper $N_{\mathrm{cit}} = C_{\mathrm{cit}} N_{\mathrm{aut}}^{p_{\mathrm{cit}}}$. The total number of citations N_{totcit} received by the papers of a collaboration then scales as

$$N_{\mathrm{totcit}} = N_{\mathrm{cit}} N_{\mathrm{pap}} = C_{\mathrm{cit}} C_{\mathrm{pap}} N_{\mathrm{aut}}^{p_{\mathrm{totcit}}}, \qquad p_{\mathrm{totcit}} = p_{\mathrm{cit}} + p_{\mathrm{pap}}. \tag{2}$$

Scaling of the total number of citations

Assuming a collective rational behaviour, and that on average the number of citations received by scientific papers may be considered as a reasonable proxy for their quality, we might expect that individual and collective choices would lead at equilibrium to

$$p_{\mathrm{totcit}} \approx 1, \tag{3}$$

namely that the total number of citations received by collaborations scales, on average, with the number of members. This expresses the fact that the (average) value of work made by N_{aut} scientists should approximately be equal to N_{aut} times the work made by a single scientist. A lower power p_{totcit} would arise in the presence of gift authorships, namely of authors who sign papers without substantially contributing.

This means that the total number of citations is not a fair indicator for authors, as it grows with the number of co-authors. Similarly, the *h*-index introduced by Hirsch (2005), being on average proportional to square root of the number of citations, grows on average as the square root of the number of co-authors.

A bibliometric index which does not overestimate nor underestimate individual contributions to a collaboration is then the number of fractionally-counted citations N_{fcit} received by each author. This means that a fraction f_A of each paper is attributed to each co-author A such that the fractions f_A sum up to unity. [1] On average N_{fcit} scales with power index $p_{\mathrm{fcit}} = p_{\mathrm{totcit}} - 1 \approx 0$, showing that N_{fcit} is a scale-invariant quantity that (unlike the number of citations) cannot be arbitrarily inflated grouping authors.

Scaling of the total number of papers

In order to implement these concepts into actual bibliometric indices for the total number of papers or for the average number of citations, we must offer arguments in favour of explicit values for the exponents p_{pap} and p_{cit}. [2]

We present here a simple "theoretical" argument. As the goal of collaborations is achieving more than what single authors can achieve, we expect $p_{\mathrm{cit}} > 0$. Assuming rational behaviour

[1] We do not address the relative assignment of credit. In some fields the contribution of different authors is reflected by their order, with special recognition given to first and last authors. Various proposals have been put forward to encode the relative credit in the fractions f_A. In some other fields authors are sorted alphabetically, giving no information about who contributed more. This happens in most papers in our data-base, so that we will assume a common f_A equal to the inverse of the number of authors.

[2] Some authors think that counting publications has no bibliometric interest, with citations being the only relevant quantity to be measured. From such a perspective, it remains nevertheless interesting to know how collaborations tend to split their bibliometric output within their publications. At the opposite extremum other authors think that citations can be distorted by social biases, and view publications numbers as a more objective bibliometric indicator.

in the formation of collaborations one may expect that (at least for not-too-big groups), individual competence of partners be as far as possible complementary, and therefore "orthogonal" in some abstract N-dimensional "space of competencies". We may therefore regard the qualitative output of a collaboration as the vector sum of N orthogonal vectors.
The limit where all authors have "orthogonal" competencies corresponds to $N = N_{\mathrm{aut}}$: then the length of such vector scales as $N_{\mathrm{aut}}^{1/2}$.[3] This corresponds to

$$p_{\mathrm{pap}} = 1/2, \qquad p_{\mathrm{cit}} = 1/2. \tag{4}$$

Sometimes, more than one collaborator is needed to fulfil a needed competency: realistic collaborations organise in $N_{\mathrm{sub}} \leq N_{\mathrm{aut}}$ sub-collaborations that work on "orthogonal" topics. We assume the number of sub-collaborations satisfies the scaling law

$$N_{\mathrm{sub}} = N_{\mathrm{aut}}^{s}, \qquad \text{with exponent } s \leq 1. \tag{5}$$

Then, the average number of citations of each paper scales as the square root of the number of "orthogonal" competencies $N = N_{\mathrm{sub}}$:

$$N_{\mathrm{cit}} \propto \sqrt{N} \propto N_{\mathrm{aut}}^{s/2}. \tag{6}$$

Assuming again an optimal distribution of resources, N_{totcit} is expected to scale as N_{aut}, and thereby the number of papers is expected to scale as

$$N_{\mathrm{pap}} \propto N_{\mathrm{aut}}^{1-s/2}. \tag{7}$$

It is reasonable to assume that the number of papers scales as N, the number of topics about which the collaboration has competencies. This leads to $s = 1 - s/2$, solved by $s = 2/3$, and thereby to

$$p_{\mathrm{pap}} = 2/3, \qquad p_{\mathrm{cit}} = 1/3. \tag{8}$$

A weaker growth of the number of papers with N leads to smaller p_{pap}.

Data about collaborations in fundamental physics

In this section we present bibliometric data in fundamental physics, that offer support for

$$p_{\mathrm{pap}} \approx 0.5 - 0.6, \qquad p_{\mathrm{cit}} \approx 0.4 - 0.5, \qquad p_{\mathrm{totcits}} \approx 1, \qquad p_{\mathrm{fcit}} \approx 0. \tag{9}$$

We use the InSpire[4] database that gives a picture of fundamental physics world-wide from ~ 1970 to 2019: 1.3 millions of scientific papers, 32 millions of references, 70 thousands of identified authors. Fundamental physics contains large collaborations, up to 3000 authors that produced 6000 publications. Adopting full counting these are counted as 3000 × 6000 publications, dominating the whole database and producing bibliometric indices uncorrelated to human evaluations of scientific merit. Fundamental physics thereby is a good sample to study

[3] Assuming that authors have different skills, the average length squared of such vector scales as N_{aut}.

[4] High-Energy Physics Literature InSpire Database (https://inspire.net/).

how the bibliometric outputs of collaborations scale with the number of collaborators. We will show data for two different kinds of collaborations:

1. **Official collaborations**. We consider the 5965 (mostly experimental) collaborations listed in the InSpire database. Each collaboration produced a certain number $N_{\rm pap}$ of papers, roughly written with the same group of $N_{\rm aut}$ authors. The left panel of fig. 1 gives some demographic information. We will show scatter plots where each collaboration is plotted as a dot, with main collaborations indicated by their name. Furthermore, we show the mean (median) as a red (magenta) curve. A blue dotted line highlights the scaling with the number of authors. [5]

2. **Occasional collaborations**. Many more multi-authored papers have been written by collaborations that form for one or few papers. To study them we proceed as follows. For each author in the InSpire database we compute the average number of authors of his/her papers, $\langle N_{\rm aut}\rangle \geq 1$, as well as his/her bibliometric indices (number of papers, of citations, etc). In view of the large number of authors we avoid showing scatter plots: we only show averages.

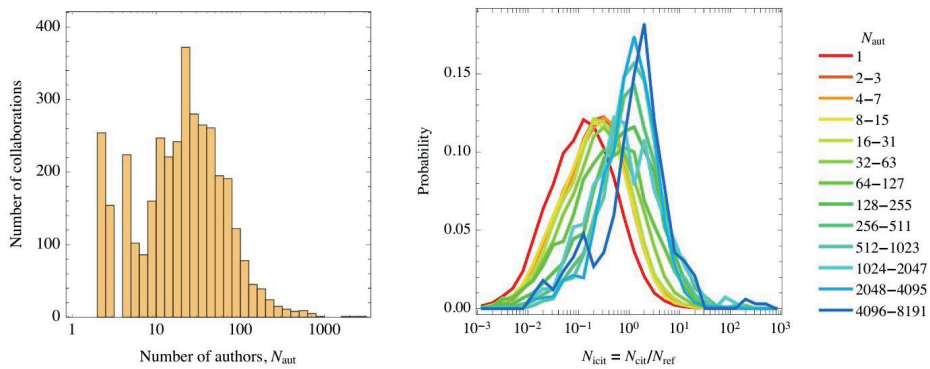


Figure 1. <u>Left</u>: Total number of collaborations listed in the InSpire database with the number of authors shown on the horizontal axis. <u>Right</u>: distribution of the number of individual citations (citations divided by the number of references of the citing papers) received by papers with the indicated number of authors.

Scaling of the total number of papers

Figure 2 shows that the number of papers produced by official (left panel) or occasional (right panel) collaborations scales with the number of authors as

$$N_{\rm pap} \propto N_{\rm aut}^{0.5-0.6}. \tag{10}$$

[5] A few collaborations varied significantly their number of authors. We define the number of authors of a collaboration by averaging the number of authors of all its papers, with weights proportional to their number of citations. This procedure assigns minor weight to proceedings written by one or few authors and to papers written by earlier incomplete phases of the collaboration.

In the right panel, theoretical papers with many authors fall below the scaling. These are rare outliers: almost all papers in theoretical categories have few authors. Theoretical papers with many authors mostly are collections of separate contributions grouped together, rather than big collaborations.

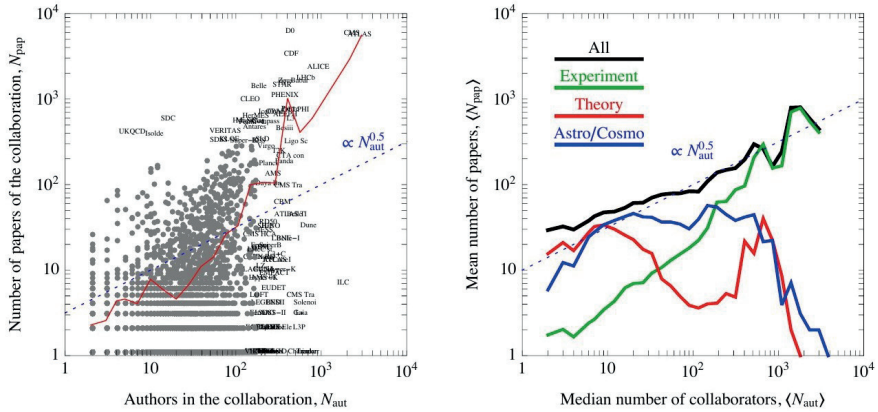


Figure 2. Number of papers versus number of collaborators.

Scaling of the mean number of citations

Figure 3 shows that the mean number of citations received by papers written by an official collaboration (left panel) or by an author (right panel) roughly scales with the average number of co-authors as

$$\frac{N_{\text{cit}}}{N_{\text{pap}}} \propto N_{\text{aut}}^{0.4-0.5}. \tag{11}$$

This result is in accordance with Hsu and Huang (2011), which (in a much smaller sample) found a power $\approx 1/3$ up to $N_{\text{aut}} \sim 10$.

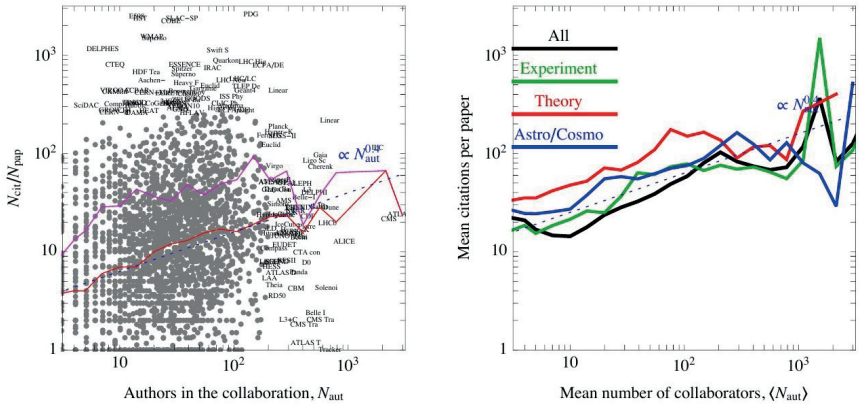


Figure 3. Mean number of citations per paper versus number of collaborators.

Scaling of the total number of citations

Figure 4 shows that the total number of citations received by an official collaboration (left panel) or author (right panel) grows roughly linearly with the average number of co-authors:

$$N_{\mathrm{cit}} \propto N_{\mathrm{aut}}^{1}. \tag{12}$$

This is expected by combining the two previous scalings: the total number of citations of a collaboration can be decomposed as the product of the number of papers written by the collaboration, times the average number of citations received per paper: these factors roughly scale as $N_{\mathrm{aut}}^{0.5-0.6} \times N_{\mathrm{aut}}^{0.4-0.5}$.

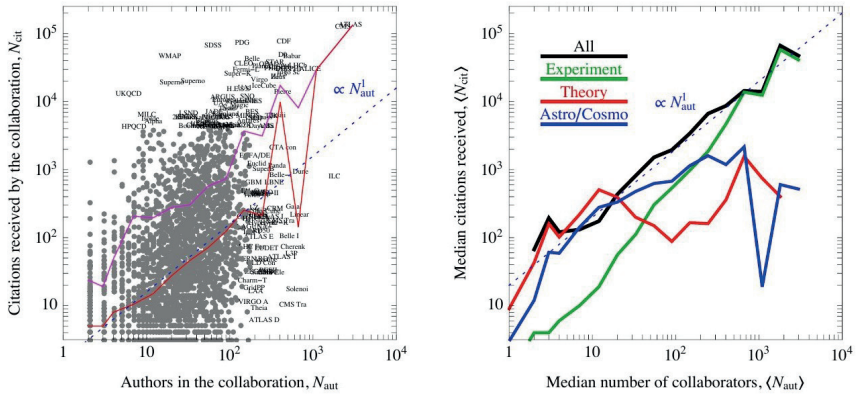


Figure 4. Number of citations versus number of collaborators.

Figure 5 shows that the total number of fractionally-counted citations $N_{\mathrm{fcit}} = \sum_p N_{p\mathrm{cit}} / N_{p\mathrm{aut}}$ received by papers p written by an official collaboration (left panel) or author (middle panel) is roughy independent of the average number of co-authors. A similar result holds for a related quantity, "individual citations", defined as fractionally counted citations divided by the number of references of the citing papers:

$$N_{\mathrm{fcit}}, N_{\mathrm{icit}} \propto N_{\mathrm{aut}}^{0}. \tag{13}$$

This means that fractionally-counted citations or individual citations neither reward nor penalise working in big collaborations, while citations or the *h*-index reward authors who prefer working in big collaborations.

Conclusions

We studied the bibliometric output of collaborations.

In the first section we presented theoretical arguments: various bibliometric indices that quantify the output of collaborations (total number of papers, of citations, etc) are expected to scale as a power of the number of authors. We presented expected values of the scaling exponents based on conservation of work, collective rational behaviour, and on plausibility argument about combining "orthogonal" competencies.

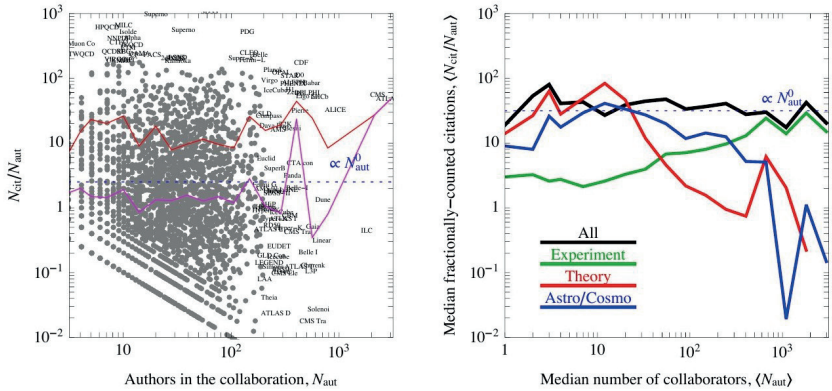


Figure 5. Total number of fractionally-counted citations versus number of collaborators.

In the second section we computed the mean and median bibliometric output of official and occasional collaborations in fundamental physics, finding that averages satisfy scaling behaviours with power indices in rough agreement with the expectations discussed in the first section.

So far we studied the means of distributions. To conclude, we extend the discussion to the full distributions, or at least to their variances. The right panel of fig. 1 shows the distributions of individual citations received by all papers in our database, splitting them according to their number of authors. We again see that papers with more authors are more cited. We also see that distributions have large variabilities: the distributions are approximatively log-normal with log-scale means that scale as $\langle N_{\mathrm{cit}}\rangle \propto N_{\mathrm{aut}}^{0.5}$ and with log-scale widths that remain approximatively constant. This behaviour is obtained from our initial theoretical considerations adding one extra assumption: that collaborations tend to equalise the total amount of skill within each competence, such that the distribution in N_{cit} of a collaboration is simply obtained rescaling the distribution of single-author papers. The bibliometric output of collaborations formed as random groups of authors would instead show larger variabilities.

Our study was motivated by the present situation in fundamental physics, where collaborations can be so large that accounting for their size has a huge bibliometric impact. The ideas that we discussed also apply to fields with smaller collaborations, up to the quantitative difference that the factors we considered have less numerical relevance.

References

Aksnes, D. W., Schneider, J. W., and Gunnarsson, M. (2012). Ranking national research systems by citation indicators. A comparative analysis using whole and fractionalised counting methods. *J. Informetrics*, *6*(1), 36-43. doi: [10.1016/j.joi.2011.08.002]

Avkiran, N. K. (2013). An empirical investigation of the influence of collaboration in finance on article impact. *Scientometrics*, *95*(3), 911-925. doi: [10.1007/s11192-012-0892-5]

Beaver, D. B. (1986). Collaboration and teamwork in physics. *Czechoslovak Journal of Physics B*, *36*(1), 14-18. doi: [10.1007/bf01599717]

Birnholtz, J. P. (2006). What does it mean to be an author? The intersection of credit, contribution, and collaboration in science. *JASIST*, *57*(13), 1758-1770. doi: [10.1002/asi.20380]

Bordons, M., Garcia-Jover, F., and Barrigon, S. (1993). Is collaboration improving research visibility? Spanish scientific output in pharmacology and pharmacy. *Research Evaluation*, *3*(1), 19-24. doi: [10.1093/rev/3.1.19]
Bouyssou, D. and Marchant, T. (2016). Ranking authors using fractional counting of citations: An axiomatic approach. *J. Informetrics*, *10*(1), 183-199. doi: [10.1016/j.joi.2015.12.006]
Carbone, V. (2011). Fractional counting of authorship to quantify scientific research output. *CoRR*, *abs/1106.0114*.
Egghe, L. (2008). Mathematical theory of the h- and g-index in case of fractional counting of authorship. *JASIST*, *59*(10), 1608-1616. doi: [10.1002/asi.20845]
Gauffriau, M. (2017). A categorization of arguments for counting methods for publication and citation indicators. *J. Informetrics*, *11*(3), 672-684. doi: [10.1016/j.joi.2017.05.009]
Gauffriau, M., Larsen, P. O., Maye, I., Roulin-Perriard, A., and von Ins, M. (2008). Comparisons of results of publication counting using different methods. *Scientometrics*, *77*(1), 147-176. doi: [10.1007/s11192-007-1934-2]
Gazni, A. and Didegah, F. (2011). Investigating different types of research collaboration and citation impact: a case study of Harvard university's publications. *Scientometrics*, *87*(2), 251-265. doi: [10.1007/s11192-011-0343-8]
Hirsch, J. E. (2005). An index to quantify an individual's scientific research output. *Proc. Nat. Acad. Sci.*, *46*. doi: [10.1007/s11192-010-0193-9]
Hooydonk, G. V. (1997). Fractional counting of multiauthored publications: Consequences for the impact of authors. *JASIS*, *48*(10), 944-945.
Hsu, J. and Huang, D. (2011). Correlation between impact and collaboration. *Scientometrics*, *86*(2), 317- 324. doi: [10.1007/s11192-010-0265-x]
Katz, J. S. and Hicks, D. (1997). How much is a collaboration worth? a calibrated bibliometric model. *Scientometrics*, *40*(3), 541-554. doi: [10.1007/BF02459299]
Lee, S. and Bozeman, B. (2005). The impact of re- search collaboration on scientific productivity. *Social Studies of Science*, *35*(5), 673-702. doi: [10.1177/0306312705052359]
Leydesdorff, L. and Bornmann, L. (2010). How fractional counting affects the impact factor: Steps towards field-independent classifications of scholarly journals and literature. *CoRR*, *abs/1007.4749*.
Leydesdorff, L. and Bornmann, L. (2011). How fractional counting of citations affects the impact factor: Normalization in terms of differences in citation potentials among fields of science. *JASIST*, *62*(2), 217-229. doi: [10.1002/asi.21450]
Leydesdorff, L. and Opthof, T. (2010). Normalization at the field level: fractional counting of citations. *J. Informetrics*, *4*, 644-646. doi: [10.1016/j.joi.2010.05.003]
Leydesdorff, L. and Park, H. W. (2017). Full and fractional counting in bibliometric networks. *J. Informetrics*, *11*(1), 117-120. doi: [10.1016/j.joi.2016.11.007]
Leydesdorff, L. and Shin, J. C. (2011). How to evaluate universities in terms of their relative citation impacts: Fractional counting of citations and the normalization of differences among disciplines. *JASIST*, *62*(6), 1146-1155. doi: [10.1002/asi.21511]
Perianes-Rodríguez, A., Waltman, L., and van Eck, N. J. (2016). Constructing bibliometric networks: A comparison between full and fractional counting. *J. Informetrics*, *10*(4), 1178-1195. doi: [10.1016/j.joi.2016.10.006]
Rousseau, R. (2014). A note on the interpolated or real-valued h-index with a generalization for fractional counting. *Aslib J. Inf. Manag.*, *66*(1), 2-12. doi: [10.1108/AJIM-01-2013-0001]
Sooryamoorthy, R. (2009). Do types of collaboration change citation? collaboration and citation patterns of south african science publications. *Scientometrics*, *81*, 177-193. doi: [10.1007/s11192-009-2126-z]
Strumia, A. and Torre, R. (2019). Biblioranking fundamental physics. *J. Informetrics*, *13*(2), 515-539. doi: [10.1016/j.joi.2019.01.011]
van Raan, A. F. J. (1997). Science as an international enterprise. *Science and Public Policy*, *24*(5), 290-300. doi: [10.1093/spp/24.5.290]
Waltman, L. and van Eck, N. J. (2015). Field-normalized citation impact indicators and the choice of an appropriate counting method. *J. Informetrics*, *9*(4), 872–894. doi: [10.1016/j.joi.2015.08.001]

Altmetrics Study of Economics

Dorte Drongstrup[1] Shafaq Malik[2] Saeed-Ul Hassan[3]

[1] *dh@bib.sdu.dk*
University Library of Southern Denmark, Campusvej 55, DK-5230 Odense M, Denmark

[2] *shafaqmalik@itu.edu.pk*
Information Technology University, 346-B, Ferozepur Road, Lahore, Pakistan

[3] *saeed-ul-hassan@itu.edu.pk*
Information Technology University, 346-B, Ferozepur Road, Lahore, Pakistan

Abstract

Social media metrics are often praised as an alternative or complement to traditional bibliometric metrics, especially in the social sciences. However, empirical investigations of the social sciences are scarce. This research in progress paper explores the extent economic research is communicated on social media platforms with an emphasis on mentions in policy documents. It examines how Altmetrics indicators relate to the Academic Journal Guide (AJG), SNIP and JCR journal rankings for economics articles. The investigated journals are selected based on Charted Association of Business School (CABS) list and extracted from Altmetric.com data. Then a statistical analysis is performed on the extracted data to get some insights into the data. The results show that the average policy count of the articles of economics journals is higher than the other subject areas mentioned in CABS and top-ranking economic journals are more likely to have articles with higher average policy count.

Introduction

During the last decade, the bibliometric measuring, evaluation, and ranking of research have become the "everyday practice" for many researchers, administrators, and funding agencies (Norris, 2019). Interestingly, due to the data availability, the recent researches also tap the contextual information (Nawaz et al., 2012) of scientific publications to leverage meta-data from full-text (Ananiadou et al., 2013). Nevertheless, this easy access to data and numerous indicators have broadened the (mis)usage of bibliometric indicators, especially at the individual level. This trend highlights the issues of lacking disciplinary coverage of the social sciences and humanities in the traditional citation databases Web of Science and Scopus. Social media metrics are often proposed as the alternative or complement to the traditional metrics (Haunschild & Bornmann, 2017) since it measures the instant usage of research publications on social media. This is often referred to as the social impact of research.

Previous studies show that the mentions and usage of research publications on social differs depending on the disciplines and platform (Hassan et al., 2017; Holmberg & Thelwall, 2014). Generally, Twitter and Mendeley show the highest usage and coverage of research publications, while it is much lower for news outlets, blogs and policy counts (Haunschild & Bornmann, 2017; Haustein, Costas, & Lariviere, 2015). Haunschild and Bornmann (2017) find less than 0.5% of all Web of Science papers have at least one policy-related mention. If divided into WoS Subject Categories, they find that Economics has the highest share (2.2%) of papers with policy mentions. Likewise, Tattersall and Carroll (2018) find that 1.4% of the University of Sheffield's papers have at least one policy mention and the majority of those papers belongs health sciences and economics. Thus, it seems that despite the low usage in many fields, the inclusion of policy mentions have the potential to show the broader impact of economics. Therefore, this paper explores the extent economic research is communicated on social media platforms with an emphasis on mentions in policy documents. Furthermore, the paper examines how Altmetrics indicators relate to the Academic Journal Guide (AJG), SNIP and JCR journal rankings for economics articles.

Data and Methods

Data is collected from Altmetric.com. This version of Altmetrics data consists of 8,158,029 JSON files, with each file representing a single publication. The AJG created by the Chartered Association of Business Schools is the starting point for the selection and analysis of articles registered in Altmetrics (CABS, 2018). The AJG list is updated every third year, first by a review board of editors and methodologists followed by a review by the scientific committee of subject experts along with, where applicable, supporting metrics (SNIP and JCR). The AJG 2018 list of journals are grouped into 22 subject areas (see table 1). These journals are ranked in five groups from 1 to 4*. Likewise, JCR and SNIP ranks are assigned to the journals based on the impact factor of each journal.

Table 1: Policy mentions across all ABS fields.

Field	Mean	N	Std. Deviation	Total Records	Percentage of total records
Accounting	1.2	283	0.602	1548	18%
Business & Economic History	1.38	265	1.119	2205	12%
Economics, Econometrics, and Statistics	2.41	23795	3.788	55560	43%
Entrepreneurship and Small Business Management	1.57	356	1.703	2075	17%
General management, Ethics, Gender, and Social Responsibility	1.34	1418	1.208	10951	13%
Finance	2.14	2546	2.911	6551	39%
Human Resource Management and Employment Studies	1.32	632	0.852	3104	20%
International Business and Area Studies	1.29	634	0.782	4957	13%
Information Management	1.2	678	0.552	9386	7%
Innovation	1.88	965	2.005	4137	23%
Management Development and Education	1.28	641	0.776	4760	14%
Marketing	1.28	650	0.891	6073	11%
Operations and Technology Management	1.17	293	0.681	2175	14%
Operations Research and Management Science	1.39	768	1.09	6761	11%
Organizational Studies	1.31	477	0.849	4294	11%
Psychology (General)	1.67	2516	1.89	15444	16%
Psychology (Organisational)	1.43	2947	1.004	11828	25%
Public Sector	1.67	2085	1.84	7790	27%
Regional Studies, Planning and Environment	1.4	1481	0.994	6337	23%
Sector Studies	1.47	3429	1.479	14823	23%
Social Sciences	1.79	5899	2.054	21442	28%
Strategy	1.32	265	0.886	1838	14%

The AJG list is a cross match based on journal title, when available, and ISSNs if journal title is not available. We created two datasets based on this cross-matching. One contains 67,775 articles published in 333 economics journals. The other contains 299,514 articles published in 1565 journals belonging to all the 22 subject areas mentioned in CABS (2018). We mapped the articles to an extensive list of Altmetrics factors, but will in this study focus on Tweet count, Policy count, Facebook count, Blog count, News count and Wikipedia count.

To obtain the source data, JCR uses approximately 11,000 journals indexed in the web of science database. The journal impact factor (JIF) formula uses articles citation counts from the

previous two years but also offers a five-year JIF. As the impact factors can be highly volatile across years, therefore, the AJG contains the JCR ranking according to the five-year JIF rather than the two-year JCR. JCR ranking in AJG18 is done by calculating a standardized impact factor that is obtained by subtracting the Journal Impact Factor from Mean Impact Factor for Subject Area and then dividing it by the Standard Deviation of Impact Factor for Subject Area (CABS, 2018).

Source Normalized Impact per Paper (SNIP) measures the impact of source titles by normalizing the citation potential in the field is based on the three-year impact factors (Waltman, van Eck, van Leeuwen, & Visser, 2013). The SNIP's main advantage is that it normalizes citations across subject areas and that it does so without relying on classifications of subject areas that in turn would create limitations (CABS, 2018). The SNIP is calculated as the number of citations given in the present year to articles in the past three years divided by the total number of articles in the past three years.

For data analysis, the rankings of AJG18, JCR, and SNIP are divided into quartiles (Q1, Q2, Q3 or Q4). Q1 indicates that the journal is in the top 25% of its subject category while Q4 indicates it is in the bottom 25% of the journals in that category. To divide the rankings into quartiles, ranges are defined for AJG18, JCR and SNIP ranks. The ranges for AJG18, JCR, and SNIP along with the quartiles are shown in Table 2.

Table 2: Quartiles Ranges for AJG, JCR, and SNIP

Rankings	Q1	Q2	Q3	Q4
AJG18	4*,4	3	2	1
JCR	1-54	55-108	109-162	163-216
SNIP	1-73	74-146	147-219	220-292

Prior to the analysis, extensive data cleaning and data processing are performed on both datasets by assigning the ranks and quartiles of AJG18, JCR and SNIP against each publication of a journal according to the rankings defined in AJG. As well as corrections of the format of the publication year and removal of detected anomalies, for example, some records in both datasets have a registered the publication year 2037, 2030, 2028, etc. We found 14 instances in the economics dataset and 25 instances in the all subject areas dataset. The records containing the year anomaly are deleted from both the datasets. Thus, the final dataset for economics journals consists of 55,560 articles and for all subject areas consists of 204,029 articles.

Results

Figure 1 shows how articles of recent years are more likely to be mentioned in altmetric.com data than the articles of the previous years as Altmetrics came into being in 2011. As well as demonstrate the broad range of research still being deemed relevant in economics. Table 1 demonstrates how policy mentions are more distinct in “Economics, Econometrics and Statistics” (43%) compared to the other fields in the AJG list (7%-39%), as well as being more frequently mentioned in policies. Thus, the results correspond to the trend found by Tattersall and Carroll (2018) and Haunschild and Bornmann (2017) with policy counts showing promise in measuring the broader impact of economics.

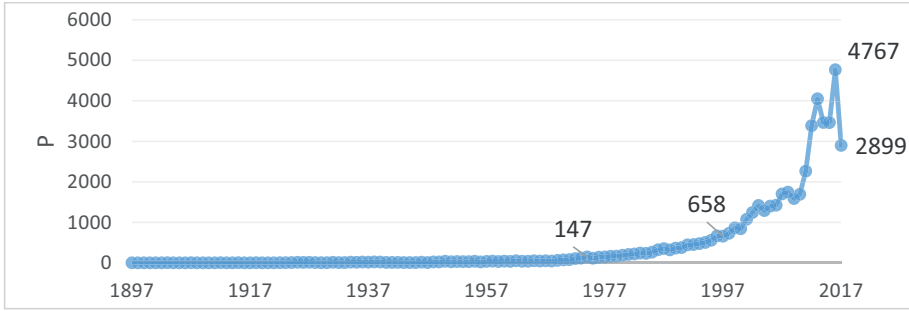


Figure 1: Publication trends of Economics related scientific literature in Altmetrics

Table 3: Descriptive data of Altmetrics mentions that exists in over 5% in the data.

	blogs	facebook	news	policy	twitter	wikipedia
Mean	1.73	1.62	3.11	2.41	5.49	1.29
N	6680	4321	3010	23,795	28,535	4809
[coverage %]	[12%]	[7.8%]	[5.4%]	[42.8%]	[51.4%]	[8.7%]
Std. Deviation	6.856	3.075	5.683	3.788	21.542	.790
Median	1.00	1.00	1.00	1.00	2.00	1.00

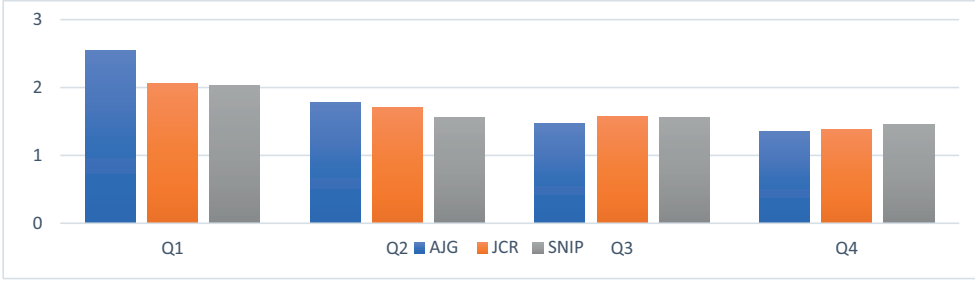


Figure 2: Average Policy mentions across the journal quality tiers

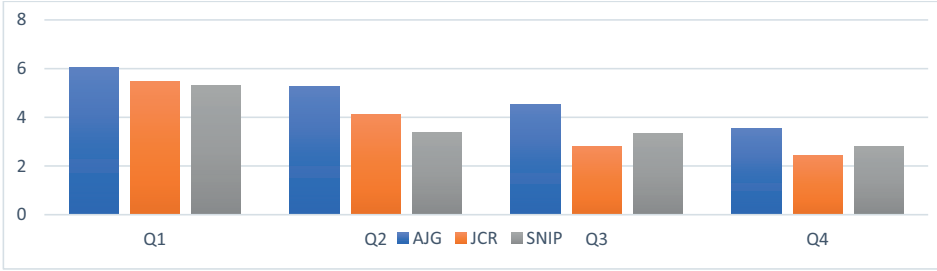


Figure 3: Average Total almetric counts across the journal quality tiers

Table 4: Spearman's rho across social mentions and journal quality tiers

Spearman's rho		AJG_Q	JCR_Q	SNIP_Q	Blogs count	Facebook count	News count	Policy count	Twitter count	Wikipedia count	Total count
AJG_Q	Correlation Coefficient	1.000	.707**	.755**	.153	-.023**	.104	.220**	.088	.073**	.111
	Sig (2-tailed)	.	.000	.000	.000	.129	.000	.000	.000	.000	.000
	N	55562	51715	55205	6674	4322	3010	23795	28538	4820	55562
JCR_Q	Correlation Coefficient	.707**	1.000	.840**	.123**	-.041	.074**	.212	.196**	.051	.195**
	Sig (2-tailed)	.000	.	.000	.000	.012	.000	.000	.000	.000	.000
	N	51715	51715	51715	6420	3863	2868	22827	26089	4634	51715
SNIP_Q	Correlation Coefficient	.755**	.840**	1.000	.107**	-.069**	.057**	.205**	.156**	.058**	.162**
	Sig (2-tailed)	.000	.000	.	.000	.000	.002	.000	.000	.000	.000
	N	55205	51715	55205	6640	4245	2988	23772	28241	4811	55205
Blogs count	Correlation Coefficient	.153**	.123**	.107**	1.000	.229**	.309**	.121**	.286**	.094**	.451**
	Sig (2-tailed)	.000	.000	.000	.	.000	.000	.000	.000	.022	.000
	N	6674	6420	6640	6680	768	1451	1959	3444	591	6674
Facebook count	Correlation Coefficient	-.023	-.041*	-.069**	.229	1.000	.215	.030*	.351	.102*	.415
	Sig (2-tailed)	.129	.012	.000	.000	.	.000	.512	.000	.155	.000
	N	4322	3863	4245	768	4322	490	497	3141	196	4322
News count	Correlation Coefficient	.104**	.074**	.057**	.309**	.215**	1.000	.049**	.203**	.015**	.483**
	Sig (2-tailed)	.000	.000	.002	.000	.000	.	.108	.000	.780	.000
	N	3010	2868	2988	1451	490	3010	1055	1935	327	3010
Policy count	Correlation Coefficient	.220**	.212**	.205**	.121**	.030**	.049**	1.000	.019**	.178**	.804**
	Sig (2-tailed)	.000	.000	.000	.000	.512	.108	.	.245	.000	.000
	N	23795	22827	23772	1959	497	1055	23796	3568	1492	23795
Twitter count	Correlation Coefficient	.088**	.196**	.156**	.286**	.351**	.203**	.019**	1.000**	.047**	.885**
	Sig (2-tailed)	.000	.000	.000	.000	.000	.000	.245	.	.135	.000
	N	28538	26089	28241	3444	3141	1935	3568	28538	1011	28538
Wikipedia count	Correlation Coefficient	.073**	.051**	.058**	.094**	.102**	.015**	.178**	.047**	1.000	.422**
	Sig (2-tailed)	.000	.000	.000	.022	.155	.780	.000	.135	.	.000
	N	4820	4634	4811	591	196	327	1492	1011	4820	4820
Total count	Correlation Coefficient	.111**	.195**	.162**	.451**	.415**	.483**	.804**	.885**	.422**	1.000
	Sig (2-tailed)	.000	.000	.000	.000	.000	.000	.000	.000	.000	.
	N	55562	51715	55205	6674	4322	3010	23795	28538	4820	55562

**. Correlation is significant at the 0.01 level (2-tailed).

*. Correlation is significant at the 0.05 level (2-tailed).

Table 3 displays the prevalence of Altmetrics factors in more than 5% of the mapped economics journal articles. It is apparent that both Twitter and policy counts are critical to demonstrating the broader impact and dissemination of economics followed by blogs, Facebook, Wikipedia, and news. Furthermore, Spearman's rho is applied on the dataset to get the correlation between the journal quality tiers and all the social mentions (see table 4). The analysis finds that the publications having

high average policy mentions are more likely to fall in the first quartile (Q1) of the journal rankings (see also Figure 2). It shows that the policy mentions hold significance in predicting the quartiles for the journal rankings. Figure 3 shows that the publications having high average total Altmetric counts are also more likely to fall in the Q1 of the journal rankings. Interestingly, Facebook mentions are slightly negatively correlated with the journal rankings for Economics, so it is rarely publications from the top-journals being shared on Facebook.

Concluding remarks

This paper demonstrates how policy mentions can help show the broader impact of economics, where a large share of the mapped publications has at least one policy mention. These results correspond to findings by previous research examining policy mentions (Haunschild & Bornmann, 2017; Tattersall & Carroll, 2018). The study also confirms that research publications from economics compared to other business associated social science fields have a greater tendency to be applied in policies (see table 1). We find that the publications published in higher ranking journals have a greater tendency to be used in policies than publications from lower-ranking journals. As well as there is a positive relationship between most of the Altmetrics indicators and the journal ranking, thus only Facebook has a negative correlation with the journal rankings. The positive relationship could because of the greater visibility of the higher-ranking journals among government officials and other researchers, but it could also be other characteristics of the articles such as research focus and methods, the number of authors or authors affiliation. Therefore, the next step is to explore further whether some article characteristics could be related to greater social media visibility. In addition to exploring the relationship between Altmetrics indicators and other bibliometric indicators for economic and the other AJG fields research.

References

Ananiadou, S., Thompson, P. and Nawaz, R., 2013, March. Enhancing search: Events and their discourse context. In International Conference on Intelligent Text Processing and Computational Linguistics (pp. 318-334). Springer, Berlin, Heidelberg.

CABS. (2018). *AJG Academic Journal Guide - Methodology*. Retrieved from Chartered Association of Business Schools: https://charteredabs.org/academic-journal-guide-2018/

Hassan, S.-U., Imran, M., Gillani, U., Aljohani, N. R., Bowman, T. D., & Didegah, F. (2017). Measuring social media activity of scientific literature: an exhaustive comparison of scopus and novel altmetrics big data. *Scientometrics, 113*(2).

Haunschild, R., & Bornmann, L. (2017). How many scientific papers are mentioned in policy-related documents? An empirical investigation using Web of Science and Altmetric data. *Scientometrics, 110*(3)

Haustein, S., Costas, R., & Lariviere, V. (2015). Characterizing Social Media Metrics of Scholarly Papers: The Effect of Document Properties and Collaboration Patterns. *Plos One, 10*(3).

Holmberg, K., & Thelwall, M. (2014). Disciplinary differences in Twitter scholarly communication. Scientometrics, 101(2), 1027-1042.

Nawaz, R., Thompson, P. and Ananiadou, S., 2012. Identification of Manner in Bio-Events. In LREC (pp. 3505-3510).

Norris, S., 2019. Bibliometrics and Research Evaluation: Uses and Abuses. Journal of Librarianship and Scholarly Communication, 7(1). DOI:http://doi.org/10.7710/2162-3309.2286

Tattersall, A., & Carroll, C. (2018). What Can Altmetric.com Tell Us About Policy Citations of Research? An Analysis of Altmetric.com Data for Research Articles from the University of Sheffield. 2(9).

Waltman, L., van Eck, N. J., van Leeuwen, T. N., & Visser, M. S. (2013). Some modifications to the SNIP journal impact indicator. Journal of Informetrics, 7(2), 272-285.

A Better Visualization for Mapping Science using Deep Learning

Ting Chen[1] Guopeng Li[2] Qiping Deng[3] and Xiaomei Wang[4*]

[1] *chenting@casisd.cn*
National Science Library, Chinese Academy of Sciences, 100190, Beijing (China)
University of Chinese Academy of Sciences, 100049, Beijing (China)
Institutes of Science and Development, Chinese Academy of Sciences, 100190, Beijing (China)

[2] *liguopeng@casisd.cn*
[4*] *wangxm@casisd.cn*
Institutes of Science and Development, Chinese Academy of Sciences, 100190, Beijing (China)

[3] *dengqp@uestc.edu.cn*
Library, University of Electronic Science and Technology of China, 611731, Chengdu (China)

Abstract

The visualization of bibliometric networks is an important part of science mapping. This study aims to provide a new way for visualizing the complex bibliometric relationships using deep learning embedding models. We compared the map generated by the classic force-directed method with four maps generated by different deep learning models. The qualities of visualization by different embedding algorithms are evaluated visually and metrically. Embedding based maps preserve a very similar global structure with the classic force-directed map, but the local structure has been improved significantly. Among them, the node2vec model has the best overall visualization performance.

Introduction

We now live in the era of the big data, understanding and mining the large-scale scientific network have created big opportunities for scholars, universities and funding agencies. Visualization techniques are often used in the map of science for understanding the evolution of scientific knowledge, dynamics of scientific disciplines, interdisciplinary and structure of science. Many different approaches have been proposed for visualizing the map of science, two of the most common approaches are the distance-based approach and the graph-based approach (Van Eck & Waltman, 2014). Both distance-based and graph-based approaches are very similar, the biggest difference is whether the distance between nodes indicates the similarity of nodes or not. For the distance-based approach, it is often used to depict topics and relationships of them within a scientific discipline (Griffith, Small, & et al., 1974; Boyack, Klavans & Börner, 2005; Boyack & Klavans, 2014). For the graph-based approach, it is suitable for identifying landmark works and members of specialties to become familiar with a field (Chen, 1999; White, 2003).

For large academic bibliometric networks, the most popular visualization approach is the distance-based approach via Multidimensional scaling (Kruskal, 1977; Liu, 2005) or VxOrd (Davidson, Wylie, & Boyack, 2001), which are used in two well-known visualization tools: VOS viewer (Van Eck, Waltman & et al., 2010) and Gephi (Davidson, Wylie, & Boyack, 2001). While these visualization methods and tools have many successful applications in the field of bibliometrics, but there are still some drawbacks for visualizing the large networks. One of the drawbacks is the lack of stability. For instance, the number of edges must be limited for visualizing a large network, but the network structure is very sensitive to the number of edges, and different trimming strategies will have some impacts on visualization map. In addition, the visualization methods themselves have a strong randomness effect, sometimes the visualization results are very different even we run the program with the same parameters. Another drawback is the loss of local structural information. In a good science map visualization, papers belonging

to the same topic need to be aggregated together in the map, and there should be a clear outline between the topics. At the same time, there also should be some differences within the topic, which can reflect the differences of sub-topics. Existing large-scale network visualization techniques usually have a good effect on displaying the global structure, but not for displaying the local structure. Local structure refers to the details and differences inside one research field or topic in this case. In this work, we try to introduce deep learning model into bibliometric visualization, so as to get a more detailed and stable map while processing large-scale data sets efficiently.

Data

To test our proposed visualization approach, 47,294 highly cited papers in the Essential Science Indicators (ESI) and their 3.6 million co-citation relationships were used as test data. Two reasons we use the ESI highly cited paper: The first reason is that highly cited papers are often used to build up various maps of science. Second, the ESI database also includes research fronts information, so we can utilize the research fronts as a possible standard to evaluate visualization methods.

Proposed approach

Network embedding plus t-SNE

Classic network representation simply uses an adjacency matrix to represent a graph G = (V,E). The elements (V) of the matrix indicate whether pairs of vertices (E) are adjacent or not in the graph. The large bibliometrics network represented in this way is a recessive high-dimensional and extremely complex space, it makes network analysis and visualization very difficult.

To tackle the challenge, we try to use the neural network to learn the representation of nodes in the bibliometric co-citation network. Once we can reconstruct the network in a low-dimensional (100-200 dimensional) vector space, all the machine learning and deep learning methods become applicable, including the state of art visualization algorithm t-Distributed Stochastic Neighbor Embedding(t-SNE) (Maaten & Hinton 2008). t-SNE is a nonlinear dimensionality reduction technique that is well-suited for embedding high-dimensional data into a 2D or 3D space. t-SNE is rarely applied to science mapping data, even though it is commonly used in another field, such as biological information, news text data, etc. (Li, Cerise & Yang, 2017; Pezzotti, Lelieveldt & et al., 2017; Liu, Bremer & et al., 2015).

Figure 1 is the flowchart of the visualization approach proposed. It comprises two main steps, firstly, we extract the low-dimensional features of nodes in the network by applying network embedding models, three popular network embedding models including node2vec (Grover & Leskovec, 2016), deepWalk (Perozzi, Al-Rfou & Skiena,2014), Line (Tang, Qu & et al., 2015) and a language model Doc2vec (Le & Mikolov, 2014) were tested in this study. Secondly, the t-SNE dimensionality reduction algorithm was utilized to project nodes from low-dimensional space into two dimensions map. After that, we compared our embedding visualization approach with the OpenOrd visualization based on co-citation network.

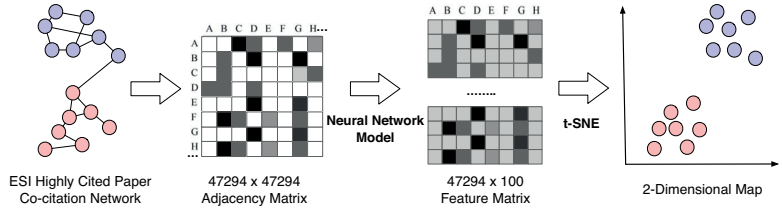


Figure 1. The flowchart of the proposed visualization approach

Base map

Figure 2 is the base map created by OpenOrd in Gephi with 47,294 highly cited paper and 3.6 million co-citation relationships. As can be seen from the figure, the map made by the popular visualization tool is great at retaining the global structure, every discipline presents its own distinct area on the map. However, the local structure is poor, there are no clear boundaries between disciplines, no distinction between sub-disciplines (topics). Besides, there are many outliers distributed in the map. This refers to the poor cohesion of the map. In order to generate more meaningful visualization layout of science map, we are looking for an approach that preserves more local structure within topics and at the same time preserving the global structure as much as possible. We will use this base map to visually compare the maps which will be created by different embedding models in the next section.

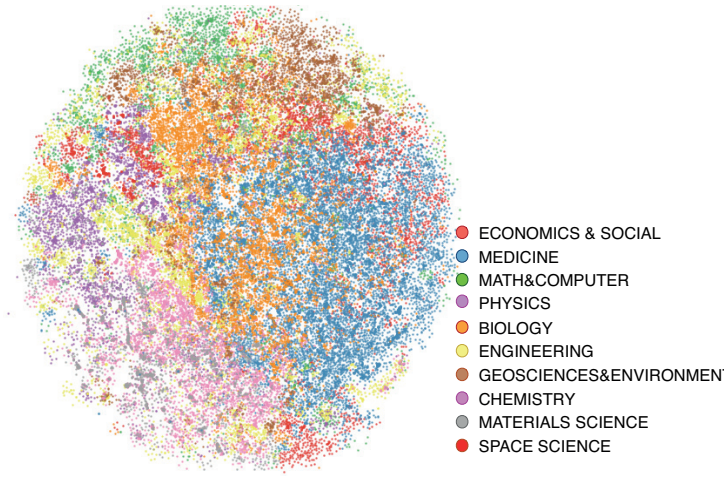


Figure 2. A base map created by OpenOrd with 47,294 highly cited paper and 3.6 million co-citation relationships

Evaluation

The evaluation of data visualization is usually subjective. Here we borrow the approach adopted by the machine learning to evaluate the clustering quantitatively. As mentioned in the previous section, the co-citation network of ESI highly cited papers was used as our test dataset, each paper has its research front information (which created by network clustering), we regard research front of each paper as the topic label of the paper. Ideally, the papers in the same research front should have stayed closer than paper from other research front in the map, and there should be a boundary between different research fronts. According to the research fronts information, some of the evaluation metrics such as Davies-Bouldin Index (Davies & Bouldin, 1979), Calinski-Harabaz Index (Caliński, & Harabasz, 1974) can be used for evaluating the performance of visualization methods. Both of the indexes are designed for measuring the variance between topics and the variance within each topic, for Calinski-Harabaz Index, higher score relates to an approach with better-defined topics. For Davies-Bouldin Index, lower score relates to an approach with better-defined topics. They are both often used as a metric for evaluating clustering algorithms, in this case, we use them as metrics.

Results

All the following results are executed on a machine with 128GB memory, 32 cores at 2.13GHz, only one core ever used in all the experiments. The running time of four embedding models are

very different, For Line model, the running time is 182 minutes, the deepWalk model takes 144 minutes, and the node2vec model is a little bit faster than Line and deepWalk, it takes 93 minutes to retrieve networking vectors. The Doc2vec model we used in this paper is a pre-trained language model based on title and abstract from the web of science indexed paper, it only takes less than a minute to retrieve all the documents' vectors, but pre-training takes hours. The process of visualization in t-SNE is also very time consuming, especially on large-scale data sets. t-SNE model takes 46 minutes to create a 2-dimension map from 100-dimension network embedded space.

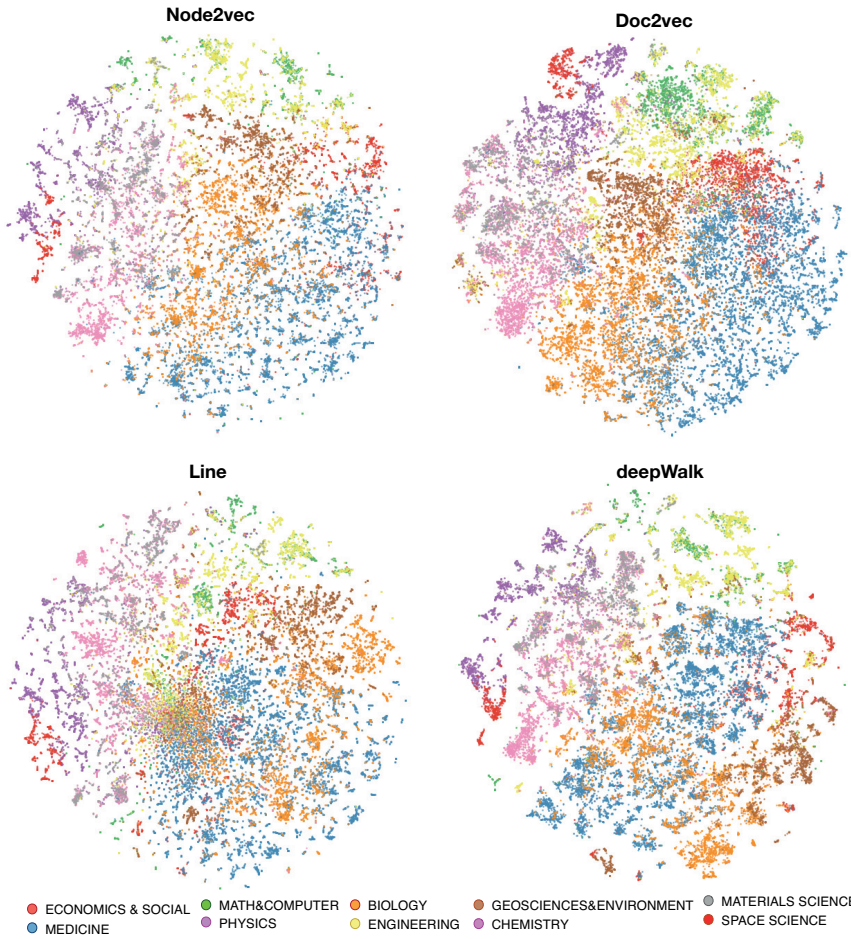


Figure 3. Co-citation network visualization of Highly cited paper by different embedding model with t-SNE

The quality of network visualization by different embedding algorithms was evaluated visually in Figure 3, each highly cited paper from ESI was mapped into a two-dimensional space as a point, distinct colour on the points represent the scientific disciplines. As can be seen, all four embedding models preserve the very similar global structure with OpenOrd (Figure 1). Medicine and Biology are placed at right bottom part of the map. Space science, Physics, Chemistry and Material are placed on the left side. Math, Computer and Engineering are placed on the top of the map. Social and business are placed on the right side beside. The Geoscience

and Environment are placed at the centre of the map. However, if we look at the local details, the embedding visualization maps are very different. All four visualization methods have some improvements in the local structure, but some models are better than others. The results of Line are not quite satisfied because the points at the centre part of the map which belonging to different disciplines are mixed together. For the deepWalk model, the small sub-topics are formed, however points from Biology, Medicine and Geoscience are still mixed a little in the centre part. The global structure of node2vec and Doc2vec map is very similar, but the visualization of node2vec performed much better in both sub-topic separation and boundary aspects. It successfully uncovers hidden local structures in the data, exposing the sub-topics in each discipline. Moreover, outliers in OpenOrd map does not appear to be outliers in node2vec map, most of these original outliers moved into the sub-topics.

The Evaluation of visualization results are presented in table 1 quantitatively demonstrate the superiority of node2vec model in visualizing the highly cited paper's co-citation relationships. The highest Calinski-Harabasz Index and the lowest Davies-Bouldin Index represent the map of the node2vec model is closer to the results of the ESI research fronts. Surprisingly the Doc2vec has second-best performance, this may prove that if there is a large amount of training data and a good model. The text features can also as good as the citation relationship. All four embedding models were tested with default parameters. For node2vec and deepWalk models, the number of walks is 10, walk length is 80, q and p are 1. For Line model, the negative ratio is 5, the order is 3. For Doc2vec model, the distributed memory was used, the window size is 5, the minimum count is 5. For the t-SNE embedding model, the perplexity is 15, early exaggeration is 30, the learning rate is 190.

Table 1. Evaluation results on 47,294 highly cited paper network visualization

	Calinski-Harabasz Index	Davies-Bouldin Index
Node2vec	**114.786**	**23.97**
deepWalk	97.22	30.77
Line	86.24	29.94
Doc2vec	103.18	28.91

Discussion and future work

We explored the applicability of network embedding model based on deep learning for visualizing the bibliometric network, and make these observations:(i)t-SNE plus deep learning network embedding models are able to separate papers from different scientific disciplines, four embedding models preserves a very similar global structure with classic OpenOrd algorithm; (ii)Visualization of node2vec model performed the best in both sub-topic separation and boundary aspects, it successfully uncovers hidden local structures in the data, exposing the sub-topics in each discipline. (iii)Unlike OpenOrd, t-SNE plus node2vec are more robust with respect to the presence of outliers.

Based on the experimental results of this paper, using the nod2vec model to embed the citation network into the feature space and then using t-SNE to display papers in two-dimensional space could be a better visualization solution for the map of science.

As this study is only at the explorative stage, we need to further test the network embedding model from different perspectives. In this paper, we only use the map of co-citation network as the base map, it should also be compared with the map of BM25 similarity network which is another very commonly used network similarity for visualizing the map of science. Furthermore,

we plan to test the stability of the proposed visualization approach by reducing the edges of the nodes when the model was trained.

Acknowledgments

This work was supported by the strategic research project of the Development Planning Bureau of the Chinese Academy of Sciences, and by the Youth Fund Project of Institutes of Science and Development, Chinese Academy of Sciences.

References

Boyack, K. W., Klavans, R., & Börner, K. (2005). Mapping the backbone of science. *Scientometrics*, 64(3), 351-374.

Boyack, K. W., & Klavans, R. (2014). Creation of a highly detailed, dynamic, global model and map of science. *Journal of the Association for Information Science and Technology*, 65(4), 670-685.

Caliński, T., & Harabasz, J. (1974). A dendrite method for cluster analysis. *Communications in Statistics-theory and Methods*, 3(1), 1-27.

Chen, C. (1999). Visualising semantic spaces and author co-citation networks in digital libraries. *Information Processing & Management*, 35(3), 401-420.

Davies, D. L., & Bouldin, D. W. (1979). A cluster separation measure. *IEEE transactions on pattern analysis and machine intelligence*, (2), 224-227.

Davidson, G. S., Wylie, B. N., & Boyack, K. W. (2001, October). Cluster stability and the use of noise in interpretation of clustering. *In IEEE Symposium on Information Visualization*, 2001. INFOVIS 2001. (pp. 23-30). San Diego: IEEE.

Eck, N. J. V., & Waltman, L. (2010). Software survey: vosviewer, a computer program for bibliometric mapping. *Scientometrics*, 84(2), 523-538.

Grover, A., & Leskovec, J. (2016). node2vec: Scalable feature learning for networks. *In Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 855-864). San Francisco: ACM.

Griffith, B. C., Small, H. G., Stonehill, J. A., & Dey, S. (1974). The structure of scientific literatures ii: toward a macro- and microstructure for science. *Science Studies*, 4(4), 339-365.

Kruskal, J. B. (1977). Multidimensional scaling and other methods for discovery structure. In K. Enslein, A. Ralston, & H. Wilf (Eds.), *Statistical methods for digital computers*. New York: Wiley.

Liu, Z. (2005). Visualizing the intellectual structure in urban studies: a journal co-citation analysis (1992-2002). *Scientometrics*, 62(3), 385-402.

Li, W. , Cerise, J. E. , Yang, Y. , & Han, H. . (2017). Application of t-sne to human genetic data. *Journal of Bioinformatics and Computational Biology*, 15(04), 1750017.

Liu, S. , Bremer, P. T. , Thiagarajan, J. J. , Srikumar, V. , & Pascucci, V. . (2017). Visual exploration of semantic relationships in neural word embeddings. *IEEE Transactions on Visualization and Computer Graphics*, PP(99), 1-1. Los Angeles: IEEE.

Le, Q., & Mikolov, T. (2014, January). Distributed representations of sentences and documents. *In 31st International Conference on Machine Learning* (pp. 1188-1196). Beijing: ICML.

Maaten, L. V. D., & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of machine learning research*, 9(Nov), 2579-2605.

Perozzi, B., Al-Rfou, R., & Skiena, S. (2014, August). Deepwalk: Online learning of social representations. *In Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 701-710). New York: ACM.

Pezzotti, N., Lelieveldt, B. P., van der Maaten, L., Höllt, T., Eisemann, E., & Vilanova, A. (2016). Approximated and user steerable tSNE for progressive visual analytics. *IEEE transactions on visualization and computer graphics*, 23(7), 1739-1752. Los Angeles: IEEE.

Tang, J., Qu, M., Wang, M., Zhang, M., Yan, J., & Mei, Q. (2015, May). Line: Large-scale information network embedding. *In Proceedings of the 24th international conference on world wide web* (pp. 1067-1077). Florence: WWW.

White, H. D. (2003). Pathfinder networks and author co-citation analysis: a remapping of paradigmatic information scientists. *Journal of the American Society for Information Science and, Technology,* 54(5), 423-434.

What affects the venture capital for start-ups: from the perspective of patent signal of Chinese bio-pharmaceutical

Lili Zhang[1], Ying Guo[2], Jun Su[3] and Cui Huang[4*]

[1] *1983803363@qq.com*
School of Public Policy and Management, Tsinghua University, Beijing 100084, China

[2] *guoying_bit@163.com*
School of Management and Economics, Beijing Institute of Technology, Beijing 100081, China

[3] *sujun@tsinghua.edu.cn*
School of Public Policy and Management, Tsinghua University, Beijing 100084, China

[4] *huangcui@zju.edu.cn, corresponding author*
School of Public Affairs, Zhejiang Univeristy, Hangzhou 310057, China

Abstract

Venture capital (VC) provides strong guarantee and financial support for enterprise innovation. As an important achievement of enterprise innovation, patents can influence VC decision-making through signal transmission. To understand this influence, we first analyse the transmission mechanism of patent signals in VC from two aspects: the patent signals that affect VC (technological signal, commercial signal, and legal signal) and other factors affecting patent signal transmission. The hypotheses are put forward on the mechanism of patent signal affecting VC, and multiple regression models are constructed to test related hypotheses. Finally, the bio-pharmaceutical industry in China is taken as an example for empirical research. The results show that the legal signal has the greatest impact on VC, followed by the technological signal, and the commercial signal has less influence. The invention patents, especially the public invention patents, have the most significant impact on VC.

Introduction

For start-ups, one of the major obstacles to securing VC is how to signal their value to potential investors. Venture capitalists tend to assess a firm by analysing considerable data on the firm's history and its perceived market potential (Useche, 2014). For an established firm, quality can be reflected through the skills of its employees, how well its products or services are designed to meet customer needs, the efficiency of its supply chain, proprietary rights, budgets, sales, income, marketing research, development activities, and so on (Zacharakis & Meyer, 2005). However, start-ups are usually young, unprofitable, and more likely to be insolvent. Therefore, the relationship between start-ups and venture capitalists is usually characterized by information asymmetry, which makes it difficult for an investor to evaluate a start-up's true potential (Baum & Silverman, 2004).

Previous studies have shown that patents have positive effect on the amount of VC financing. But it seems difficult to get a clear conclusion according to present research. What makes matters more urgent, there is almost no scholars have researched similar topics in in China, a different and more sophisticated market. Bio-pharmaceuticals are an essential contributor to China's national economy and for their role in maintaining people's health and improving their quality of life. However, this industry is also classified as high risk due to its advanced technology, high investment requirements, and long R&D cycles. This means the potential for returns is immense – a characteristic that is favoured by some investors but may turn others away. These features make the bio-pharmaceutical industry in China suitable subject for studying the impact of patent signals on VC financing.

In the analysis that follows, we first explain the types of signals a patent can convey from the start-up to the investor and their way of transmission. Based on our previous research into the nature of patents and extant theories, patents transmit three types of signals – technological,

commercial, and legal. Our hypotheses are based on specific attributes of a patent, the specific types of signals they send, and how each attribute affects the level of VC funding a start-up receives. To test our hypotheses in an empirical setting, we built a database in China's bio-pharmaceutical industry and constructed a regression model to evaluate the effect of different variables in the sample. Our conclusions should be of value to start-ups by assisting them to signal more effective signals to investors through patents. Investors can use these results to assess a start-up's potential more objectively before making a final investment decision. Both can decrease the venture in financing.

Theory and hypotheses

Several scholars have already studied how patent signals impact VC decisions. However, despite the progress made by these studies, they have some shortcomings. First, there is little consideration of a patent's quality. In analysing the impact of patents on venture capital, most studies only consider the number of patents. And, in the rare research that does include the quality of a patent, there is still much controversy about whether the same patent would have a different influence on the final investment depending on who and how it was assessed. Some patent attributes can be judged directly, while other content can only be accurately evaluated by expert intelligence. The former is more objective, while the latter is more difficult and costly to implement. Moreover, the results may vary depending on who is selected to act as an expert. Second, there is insufficient consideration of the Chinese context. The VC market in China is developing very quickly and, as with all countries, its characteristics are unique. Therefore, the conclusions derived from other markets are difficult to generalize to China.

Patents are not simply the output of a start-up's technological innovation, they are also an important input for production, management, and ongoing company operations (Hu, 2012). Further, the legal nature of a patent conveys the innovation's uniqueness and, to some extent, its difficulty to imitate. Therefore, patents can convey three types of signals about a start-up to potential investors: the level of its technology (a technological signal), its business prospects (a commercial signal), and how well its core technology is protected (a legal signal). As outlined above, the influence of each of these signals has been debated in the literature (Lemley, 2001). However, we argue that, no matter the impact, patents do send signals, and the power of each type needs to be further explored.

Figure 1 maps out the signal transmission process from a start-up to an investor using a typical communications system as a metaphor. The transmitter (the start-up) sends out encoded information S_1, S_2, S_3 (its patents) through a channel (a patent office or regulator). The message is decoded by the receiver (the investor) into its informational value O_1, O_2, O_3 (the technological, commercial, and legal signals). The informational value is then assessed, and the investor judges whether or not to act.

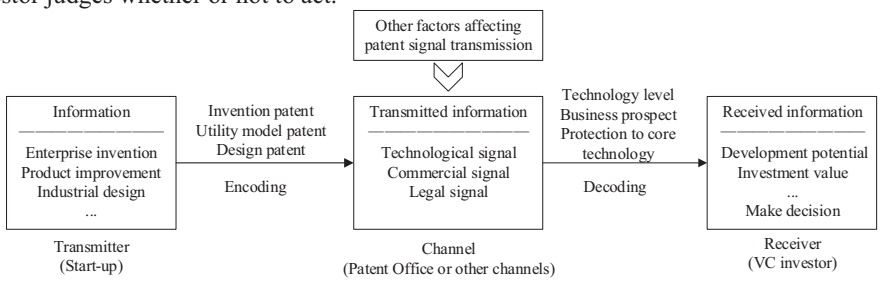


Figure 1 Signal transmission from the start-up to the investor

Based on information from the Chinese government's "Patent Search and Analysis" website (PSS)[1], we have summarized the relationship between a patent's main attributes and its technological, commercial, and legal signals in Table 1. The reason why we summarize in this way is clarified in the follow part. Note that this is a relatively simplistic summary and that the signals a patent are far more complex in reality. To make it clearer to analyse the patent signal, we consider the main signal a patent attribute transmits, which means this attribute might transmit many kinds of signals, but the investor will focus their attention on technological or commercial or legal signal in their decision on VC. The following sections describe each signal and how it is measured in detail.

Table 1 Patent attributes and their corresponding signals

Patent attribute	*Technological Signal*	*Commercial Signal*	*Legal Signal*
Status: public or granted			√
IPC classification	√		
Inventor	√		
Patentee	√	√	
Citation	√	√	
Cited	√	√	
Prosecution length	√		√
Patent age	√		√
Number of patents	√	√	
LOC classification		√	
Patent family		√	√

Technological signals

Technological innovation is the core of a patent. We hold that the corresponding attributes listed in Table 1 have the power to transmit signals to investors about the scope, degree, and maturity of a technological innovation. The IPC classification of a specific patent, and indeed the IPC classifications of all the patents held by the start-up, reflect both the scope and the depth of a start-up's technology. The inventor and patentee, the other patents cited and the citations the patent receives, and the length of prosecution length signal the degree of innovation degree. The patent's age indicates the level of technology's maturity, and the number of patents is a likely indicator of the overall level of technological innovation in that start-up.

（1） IPC classification

IPC codes are an internationally accepted method of categorizing patent documents into technological fields (Zhang et al., 2016). A common way to measure technological scope is the number of different IPCs assigned to a patent, particularly invention and utility model patents (Reitzig, 2004). The number of times a start-up has lodged a patent under a particular IPC category reflects the depth of its technology – and, to some extent its strength in that technology. Both the scope and the depth of a start-up's research need to be considered by investors, and both constitute the technological signals patents send. Therefore, we propose the following hypothesis.

H1a: The scope of a patent's technology has positive impact on the amount of VC invested.

H1b: The scope of the start-up's overall technology has positive impact on the amount of VC invested.

H1c: The depth of the start-up's research has positive impact on the amount of VC invested.

（2） Inventor and patentee

[1] http://www.pss-system.gov.cn/

The inventor and patentees play an important role in knowledge transfer. The inventor embodies the R&D capability of the firm, while the patentees reflect cooperation levels, either internally or with external partners (Lei et al., 2013). Joint patentees can also signal future business prospects or the future management structure of the start-up (Agrawal, 2006). Hence, these attributes send both technological and commercial signals about the start-up.

Patent partnerships between industry, universities, and research institutions, called "CoPatents", are fairly common. Start-ups looking for VC are generally in the early stages of development and, therefore, are highly unstable. As such, it is difficult for investors to judge the start-up's technological strength, but these reservations can be offset by collaborating with a university or research institution with a mature reputation for research and development in the field. However, investing in a joint patentee relationship may not always be viewed positively. For instance, the financier may not be able to negotiate sole rights to the technology or the joint partner may impose some unwanted constraints. Yet, overall, the "collateral" solid partners bring in terms of oversight, reputation, and future business prospects tend to work in the start-up's favour. Therefore, we propose the following hypothesis.

H2a: The number of inventors has a positive impact on the amount of VC invested.

H2b: The number of patentees has a positive impact on the amount of VC invested.

H2c: The proportion of CoPatent has a significant impact on the amount of VC invested.

（3） Citation and cited

The citations a patent receives and the other patents cited is an indicator of knowledge flow, technology transfers, originality, advancement, and the influence of one technology over another (Zhang et al., 2017). They also reflect the importance of a patent and its place in the process of technological development (Antille et al., 2005). For example, if Patent A cites B, then it is likely that Technology A is advancing Technology B by, say, incorporating a more diverse set of knowledge or progressing the technology in an original way. However, that may also mean paying royalty fees to the holders of Patent B. These two elements create an interesting juxtaposition for investors. A greater number of patent citations may mean a stronger technological advancement, but a weakened market value due to royalties and perhaps copyright constraints. The reverse is true in the case of a patent being cited. The technology may no longer be as leading-edge, but being the foundation of future technology coupled with the potential for additional revenue through royalties (Haeussler et al., 2014), has clear commercial opportunities. Therefore, patent citations and patents cited can simultaneously transfer both technological and commercial signals, which leads to the following hypothesis.

H3a: The number of patent citations has a significant impact on the amount of VC invested with transmitting positive technological signals and negative commercial signals.

H3b: The number of patents has a positive impact on the amount of VC invested.

（4） Examination length

The patent examination length is the time span between submitting a patent and a final decision. i.e., the speed of the patent examination. As a general rule, the faster the review, the higher the technological innovation of the patent (Haeussler et al., 2014). The longer the examination length, the more closely the patent is examined. Examination can take a long time if the technology is of high interest. Competitors may use public opinion or other means to prevent the patent from being publicized or authorized, or to force the applicants to be more rigorous in their justifications. Responses can be delayed to extend the application timeline. Previous studies with large data samples have shown that examination length is positively correlated to the stability of patent rights, whereas other studies have reached the opposite conclusion. We will further examine this debate and, therefore, put forward the following hypothesis.

H4: The length of the patent application process has a significant impact on the amount of VC invested.

（5） Patent age

Patent age is the length of time that has passed since the patent was filed. For a patent to remain valid, an annual fee needs to be paid to the patent office otherwise the patent may be revoked. Although there is a recovery procedure, the value of a patent is greatly reduced if an annual fee is missed because it signals that the patentee does not highly value its worth. Therefore, the older a patent the more importance a patentee attaches to the patent, i.e., the higher its technological value. However, patents in China have a period of protection. Patent Law stipulates that "the duration of invention patents is 20 years, and the duration of utility model patents and design patents is 10 years, all calculated from the date of filing". Therefore, the older the patent, the shorter the remaining period of monopoly and the smaller its legal value. As such, patent age can convey technological signals and legal signals. The current relationship between patent age and patent value is still inconclusive. Sapsalis (2006) draws a significant positive correlation between the two in a Belgian biotechnology industry study, while Gambardella et al. finds no significance (Gambardella et al, 2011). To further investigate any correlation, we make the following hypothesis.

H5: The patent age has a significant impact on the amount of VC invested with transmitting positive technological signal and negative legal signal.

（6） Number of patents

As the most basic attribute of a start-up's patents, the number of patents is a comprehensive reflection of the technological and commercial worth of the firm. However, whether the number of patents has a positive or a negative influence on VC is still inconclusive. Since patents include many unimportant innovations, they are not necessarily a reflection of the level of innovation (Gayle, 2001). In addition, the value distribution of patents is highly asymmetric. For example, Scherer and Harhoff (2004) find that 10% of patents in Germany and the US hold more than 80% of the value of all patents. Thus, the value of patents and the number of patents are not necessarily positively correlated. In Western property rights markets, the market economy system and the legal environment have been relatively perfect, while China's patent technology has a low level of marketization and insufficient intellectual property protection. This leads to the following hypothesis

H6: The number of patents has a significant impact on the amount of VC invested.

Commercial signals

Commercial signals encompass the potential for commercialization, industrialization, and marketization. These benefits represent the real value patents can bring to a start-up by converting a theoretical innovation into a practical application, taking market competition into account. The previous section outlines that the patentee, patent citations, cited patents, and the number of patents can all convey commercial signals. In addition, the LOC classification number for design patents and family patents can also send commercial signals.

（1） LOC classification

The Locarno classification system (LOC) was established in the Locarno Agreement (1968) as an international classification for registering industrial designs. Products with the same purpose or for single use only have one LOC. However, a product with two or more uses or combined uses have multiple LOCs. The number of LOCs represents the quantity of the product, which affects the size of the product market – a commercial signal. Hence, we propose the following:

H7: The number of LOC classifications has a positive impact on the amount of VC invested.

（2） Patent families

Because patents are territorial, which means their legal effects are only valid within their sovereign jurisdiction, and the patent approval system implemented in most countries is "early public, delayed review", many patents are actually a group of patents registered in different countries with the same or similar attributes, i.e, a patentee apply for a patent through Patent Cooperation Treaty (PCT), this patent can get a patent family. Hence, patent families can be

used to understand the status of individual patent applications in different countries or with international patent organizations. Moreover, they hold a wealth of information about which countries have the potential for market expansion depending on whether rights protection is underway or already granted. As such, a family of patents transmits commercial and legal signals. Based on the number of patents within a family held by a start-up, we propose the following hypothesis.

H8: The number of patents in the same family has a positive impact on the amount of VC invested.

Legal signals

The law gives patentees exclusive rights over a technology, but the strength of that exclusivity depends on both the period of protection and the stability, scope, and legal status of the protection. These are all legal signals to investors, along with the patent age, examination length, and number of patents in a family as discussed in the previous sections. As mentioned, we have limited our analysis of legal signals to invention patents. This is because invention patents in China have two legal statuses – public and granted. Granted invention patents allow the patentee to impose a monopoly within the scope permitted by law and the rights to all the income derived from the technology. Public invention patents are essentially in a six-month review period, where the details of the patent are made public and any who choose can challenge the patent's validity or ownership. The status is usually imposed if an innovation has previously been disclosed, say at an exhibition or academic conference. Studies have shown that the legal status of a patent affects VC decisions for start-ups. For example, studies on countries outside China find that public invention patents are a more accurate reflection of a start-up's technological level than granted invention patents (Ernst, 2001). These authors believe that whether or not a patent is granted is greatly influenced by human factors. However, at present, there is little evidence that the same conclusions apply to China. Therefore, to further verify the impact of an invention patent's legal status of invention patents on venture capital, we put forward the following hypothesis.

H9: The number of public and granted invention patents held by a start-up have different impacts on the amount of VC invested.

Data and methodology

Data

The start-up VC data was sourced from the Zero2IPO database, a comprehensive, accurate and timely professional database covering VC events since 1992 in China[2]. We got the data on 3 November 2017 using the industry classifications 'biotechnology' and 'medicine' as the search criteria. We used Tianyancha[3] and Qichacha[4] to collect other basic information about each start-up. We excluded any records with incomplete fields, resulting in a total of 535 VC investments into 457 start-ups (some start-ups received multiple investments). Patent data was retrieved from PSS. Retrieved patents were only recorded and counted if the patent was granted prior to a VC investment. What's more, there might be difference between out-licensed and in-licensed patents in the view of investors. So we just chose the start-ups whose patents have no record of patent transfer. Of the 457 start-ups, 227 held existing patents at the time of financing and these firms received 275 investments. However, there are only two VC events are invented by state-owned VC, the Sdicfund. It's too few to control this variable. So we delete these two events to

[2] See the detail from http://www.zero2ipo.com.cn/service/6/
[3] https://www.tianyancha.com/
[4] http://www.qichacha.com/

make sure all the VC investors are private VC, which means we get 226 start-ups and 273 events with patents.

Variables

"VC amount" is the dependent variable. However, given that different currencies cannot be compared, we converted all investments into CNY. Additionally, the dataset covers a very broad time span (1996 to 2017) and inflation rates made it impossible to directly compare the amount of funding across years. Therefore, we standardized the amount of funding.

As outlined above, China's patents are divided into three overarching types and four if the two legal status of an invention patent are counted. Each type has different attributes with different levels of importance to investors. Given the focus of this analysis, the influence a patent's signals have on financing decisions, patent attributes formed the basis of the independent variables. Further, given each attribute sends a different signal to investors, we developed separate calculations for each of the variables. Unless otherwise specified, the same method of calculating each variable was used for all three types of patents. The independent variables and their calculation methods are listed in Table 2.

Table 2. Independent variables and their calculation methods

Independent variables	*Calculation method*
Patent technology scope	$N(PatentIPC)=\sum_{i=1}^{n}N(PatentIPC)_i \Big/ n, 1\le i\le n$
Start-up technology scope	$N(StartupIPC)$
Start-up research depth	$N(IPCDepth)=\sum_{i=1}^{n}N(PatentIPC)_i \Big/ N(StartupIPC)$
Number of inventors	$N(Inventor)=\sum_{i=1}^{n}N(Inventor)_i \Big/ n$
Number of patentees	$N(Patentee)=\sum_{i=1}^{n}N(Patentee)_i \Big/ n$
Proportion of CoPatent	$r(CoPatent)=N(CoPatent)/N\_k, k\in\{INVP,\ INVG,\ UTMP,\ DSGP\}$
Number of patent citations	$N(PatentCitation)=\sum_{i=1}^{n}N(PatentCitation)_i \Big/ n$
Number of patents cited	$N(PatentCited)=\sum_{i=1}^{n}N(PatentCited)_i \Big/ n$
Prosecution length	$Length_i=DatePublic\left(Grant\right)_i-DateApply\left(Grant\right)_i, Length=\sum_{i=1}^{n}Length_i \Big/ n$
Patent age	$AgePatent_i=\left(DateVC_i-DateApply_i\right)/365, AgePatent=\sum_{i=1}^{n}AgePatent_i \Big/ n$
Number of patents	N
Number of LOCs	$N(LOC)=\sum_{i=1}^{n}N(LOC)_i \Big/ n$
Number of family patents	$N(PatentFamily)=\sum_{i=1}^{n}N(PatentFamily)_i \Big/ n$

The other factors that may interfere with the signals a patent sends were the control variables. These are defined as follows: *VCRound* (VC round) includes the following rounds: angel, seed, Pre-A, A~G, and the New Third Board round. These rounds are denoted as 1~11; *AgeStartup* (Start-up age) is a measure of the age of the start-up as at the date of the VC investment, calculated as the difference between the investment date and the establishment date of the start-up; *FirstCity* (Dummy first-tier cities) A value of 1 denotes companies located in China's first-

tier cities, such as Beijing, Shanghai, Shenzhen, and Guangzhou, and 0 otherwise; *VCForm* (Dummy investment form) *VCForm*=1 denotes an individual investment and *VCForm*=0 denotes a joint investment; *Currency* (Dummy investment currency) *Currency*=1 indicates the investment was made in CNY; *Currency*=0 means a foreign currency.

Model

We finally selected Multiple Linear Regression (MLP) model by comparing the deterministic coefficients R^2 of the multivariate linear model and the nonlinear model (i.e., the goodness of explain and fit). The following model was developed as a basic tool for analyzing the influence of patent signals on VC financing.

$$\ln(AmountVC) = \alpha_0 + \alpha_1 TechSig + \alpha_2 ComSig + \alpha_3 LegSig + \alpha_4 VCRound + \alpha_5 AgeStartup + \alpha_6 FirstCity + \alpha_7 VCForm + \alpha_8 Currency + u$$

where u is stochastic disturbance term, and *TechSig, ComSig,* and *LegSig* represent the technological, commercial, and legal signals, respectively.

However, as described in Section 2, different types or status of patents will jointly transmit these three kinds of signals. Hence, the final model used was

$$\begin{aligned}\ln(AmountVC) = {} & \alpha_0 + \beta_1 N(StartupIPC) + \beta_2 N(IPCDepth) + \beta_3 N(PatentIPC) + \beta_4 N(Inventor) \\ & + \beta_5 N(Patentee) + \beta_6 r(CoPatent) + \beta_7 N(PatentCitation) + \beta_8 N(PatentCited) \\ & + \beta_9 Length + \beta_{10} AgePatent + \beta_{11} N + \beta_{12} N(LOC) + \beta_{13} N(PatentFamily) \\ & + \alpha_4 VCRound + \alpha_5 AgeStartup + \alpha_6 FirstCity + \alpha_7 VCForm + \alpha_8 Currency + u\end{aligned}$$

Results and discussions

In our original results of linear regression, the multicollinearity of *N* and *N(StartupIPC)* are too strong (VIF>10). Here we choose Stepwise Regression to compare the R^2 and final we delete the variable *N(StartupIPC)*, whose influence is not prominent and significant, to overcome the multicollinearity. Table 3 shows the coefficients and t-values (in brackets) of the regression analysis for the four different types of patents. The final regression analysis appears in Table 4. Note that there were only a small number of utility model patents being cited, so we removed the variable *N(PatentCited)* for this type. Additionally, *N* and *N(LOC)*, were removed from all types because their multicollinearity (value of VIF) was higher. The "H" column represents the hypotheses in Section 2; "+" means we predicted the corresponding variable would be positively correlated with the dependent variable, "-" means a negative correlation, and "Δ" means the variable has a significant, yet hypothetical, influence on the dependent variable.

The regression analysis reveals some interesting results, which are explained in the following four subsections.

Table 3 Regression results of different types or status of patents

Dependent variable ln(AmountVC)	*Invention patents*		*Utility model patents*	*Design patents*	*H*
	Public	*Granted*			
N(PatentIPC)	-0.022 (-0.343)	0.111 (1.636)	-0.277*** (-3.146)		+
N(IPCDepth)	0.054 (0.842)	-0.004 (-0.068)	0.068 (0.740)		+
N(Inventor)	-0.0003 (-0.005)	0.017 (0.262)	-0.047 (-0.526)		+
N(Patentee)	0.270*** (3.010)	0.044 (0.593)	0.089 (1.074)		△
r(CoPatent)	-0.262*** (-3.052)	-0.138* (-1.907)	-0.056 (-0.656)		+

N(PatentCitation)	-0.167** (-2.448)	-0.020 (-0.311)	0.214 (0.958)		△
N(PatentCited)	0.126* (1.850)	0.088 (1.337)	/		+
Length	-0.023 (-0.309)	0.068 (1.000)	0.133 (1.400)	0.001 (0.010)	△
AgePatent	-0.028 (-0.384)	0.002 (0.023)	-0.189* (-1.886)	0.039 (0.308)	△
N	0.120* (1.915)	0.122 (1.967)	-0.059 (-0.263)	0.199* (1.913)	△
N(LOC)	/	/	/	0.083 (0.760)	+
N(Patentfamily)	0.126* (1.785)	-0.036 (-0.562)	0.087 (1.019)	-0.025 (-0.226)	+
VCRound	0.089 (1.187)	0.057 (0.822)	0.065 (0.645)	0.100 (0.818)	-
AgeStartup	-0.028 (-0.384)	0.117* (1.730)	0.165 (1.570)	0.184 (1.344)	-
FirstCity	-0.033 (-0.513)	-0.008 (-0.124)	0.087 (0.953)	-0.024 (-0.224)	+
VCForm	-0.366*** (-5.934)	-0.468*** (-7.576)	-0.400*** (-4.752)	-0.288*** (-2.676)	-
Currency	-0.246*** (-3.830)	-0.214*** (-3.441)	-0.302*** (-3.215)	-0.266** (-2.323)	-
Sample Size	210	209	103	86	
R^2	0.315	0.336	0.431	0.273	

***, **, * represent 1%, 5%, 10% significance levels. The same below.

Table 4 Regression results of all patents

Variable	*Coefficient (t score)*	*H*	*Variable*	*Coefficient (t score)*	*H*
N(PatentIPC_INVP)	-0.077 (-0.947)		*Length_INVP*	-0.050 (-0.618)	
N(PatentIPC_INVG)	0.144* (1.893)	+	*Length_INVG*	-0.119 (-1.106)	Δ
N(PatentIPC_UTMP)	-0.344*** (-3.493)		*Length_UTMP*	-0.031 (-0.321)	
N(IPCDepth)	-0.002 (-0.03)	+	*Length_DSGP*	0.237* (1.844)	
N(Inventor_INVP)	-0.085 (-0.857)		*AgePatent_INVP*	-0.103 (-1.091)	
N(Inventor_INVG)	0.097 (1.031)	+	*AgePatent_INVG*	0.071 (0.82)	Δ
N(Inventor_UTMP)	-0.058 (-0.628)		*AgePatent_UTMP*	0.122 (1.241)	
N(Patentee_INVP)	0.161 (1.294)		*AgePatent_DSGP*	-0.125 (-1.566)	
N(Patentee_INVG)	-0.039 (-0.352)	Δ	*N(PatentFamily_INVP)*	0.090 (1.127)	Δ
N(Patentee_UTMP)	0.120		*N(PatentFamily_INVG)*	0.022	

	(0.905)			(0.238)	
r(CoPatent_INVP)	-0.046 (-0.497)		*N(PatentFamily_UTMP)*	-0.046 (-0.617)	
r(CoPatent_INVG)	-0.090 (-0.956)	+	*N(PatentFamily_DSGP)*	0.074 (1.002)	
r(CoPatent_UTMP)	-0.015 (-0.285)		*VCRound*	0.034 (0.508)	-
N(PatentCitation_INVP)	-0.187** (-2.579)		*AgeStartup*	0.103 (1.51)	-
N(PatentCitation_INVG)	0.051 (0.699)	Δ	*FirstCity*	0.023 (0.418)	+
N(PatentCitation_UTMP)	0.110** (2.054)		*VCForm*	-0.412*** (-7.816)	-
N(PatentCited_INVP)	0.140* (1.829)	+	*Currency*	-0.271*** (-4.825)	-
N(PatentCited_INVG)	-0.019 (-0.265)		*Sample size*	273	
R^2			0.405		

As a whole, there aren't many variables impacting the VC decision significantly, which might because of the shortness of our samples. Nevertheless, we can still get some meaningful conclusions, even not all kinds of patents are significantly correlated.

First, the results about the scope of the technology in the public invention patents, the number of patents cited, and the number of family patents test the hypothesis. We explain the reason in the part of the hypothesis and here we don't repeat them.

Second, some results about the number of patentees, the number of patent citations, the examination length, the patent's age, and the number of patents answer the hypothesis. As we described and analysed in the second part of this paper, if the regression results show positive impact, this might because the positive impact is stronger than the negative impact. And vice versa. It's noticed about the number of patent citations, which of public invention patents and utility model patents have different impacts on VC. This might because the public invention patents will transmit more technological signals than the utility model patents.

Third, there are also some variables rejected the hypothesis. The results about the scope of the technology in the utility model patent rejected the hypothesis which might because the VC investor will think a utility model patent has no focused domain to expand the market efficiently. Also, the regression results for the ratio of CoPatents are contrary to the hypothesis. A possible explanation for this may be CoPatents imply some level of constraint. The patentees must cooperate, which has advantages and disadvantages but usually involves compromise on both sides. Additionally, collaboration in our sample was limited and usually involved a university. While, these types of partnerships can lend credibility to a start-up, they can also send a signal to investors that the start-up's "independent R&D strength is insufficient".

And forth, the depth of research, the number of inventors, the number of LOCs had no significant impact.

Overall, the signals a patent sends have a significant impact on the amount of VC funding received. But each type of signal has a different degree of influence. All the legal indicators are influential, and technological signals have a greater impact than commercial signals. These results suggest that the most important concern for investors is a patent's legal status, i.e., the scope of its rights of protection, closely followed by the level of technological innovation. For an investor, these two factors outweigh the start-up's potential for commercializing the innovation. Public invention patents have the most significant impact. These types of patents tend to reflect a higher level of technological innovation and usually represent a start-up's core

technology. However, their scope of protection is still to be determined, so investors must be cautious. China does not always grant a public invention patent, and some companies use this type of patent to artificially inflate their technological portfolio. Hence, investors are right to have doubts. The above regression results also show that invention patents do reflect innovation in core technologies and that the practicality of utility model patents can play a significant positive role in VC financing.

Conclusions

For each type of patent, different forms of investment and investment currencies affect the role patent signals play in VC decisions. After further analysing the reasons for this phenomenon, we draw the following conclusions and relevant policy suggestions.

The patent attributes that influence financing decisions can be divided into two gradients of impact intensity. The first gradient, which tends to reflect the macro factors of the start-up, always influences the level of VC investment and includes the scope of the start-up's technology portfolio, the number of patentees, and the number of CoPatents and number of patents. The second gradient, i.e., the details of a specific patent, sometimes do and sometimes do not have an influence. It depends on the types of patent. These attributes include the scope of the technology in the patent, number of patent citations, number of patents cited, patent age, examination length, number of family patents. However, the influence of all of these attributes are affected by the form of investment and the investment currency.

Our results indicate that investors pay less attention to the details in a patent and more to the macro factors of the start-up. However, some research on other countries and contexts has shown that patent details, such as citations and patents cited, do affect VC decisions. This may be because the VC market in China is still immature, so investors do not have the tools to fully analyse a patent's value when assessing a start-up. Additionally, the patent approval process in China needs to be improved so that more information is disclosed. For example, the requirements surrounding patent citations and patents cited need to be perfected in both the process of writing patent applications and the examination subsequently conducted by the patent office. Information about patents also needs to be updated and maintained in a more timely manner so that patents have more value as a reference tool.

Further, the form of investment and the currency have an impact on how effective a patent's signals are. For example, joint investors are better able to offset high risk and make up for shortfalls in funding. They are often larger and have more investment experience than a single investor. Hence, they may be more inclined to undertake a well-founded assessment of the start-up using quantifiable, objective indicators, such as R&D investment, patent level, etc. and make their decisions based on a more mature, market-oriented approach. For these investors, what may seem like risky decisions to a single investor become more sensible and reasonable.

Acknowledgements

This paper was supported by Key International Cooperative Research Project of National Natural Science Foundation of China (71520107005), Innovation Research Group Project of National Natural Science Foundation of China (71721002), Outstanding Youth Science Fund Project of National Natural Science Foundation of China (71722002), General Project of National Natural Science Foundation of China (71673164), and Independent Research Project of Tsinghua University (20165080054).

References

Agrawal, A. (2006). Engaging the inventor: exploring licensing strategies for university inventions and the role of latent knowledge. *Strategic Management Journal*, 27(1), 63-79.

Antille, J., Baumann, S., Bernasconi, A., Breedon, R. E., Camilleri, L., & Cool, R. L., et al. (2005). Market value and patent citations. *Rand Journal of Economics*, 36(1), 16-38.
Baum, J. A. C., & Silverman, B. S. (2004). Picking winners or building them? alliance, intellectual, and human capital as selection criteria in venture financing and performance of biotechnology startups. *Journal of Business Venturing*, 19(3), 411-436.
Ernst, H. (2001). Patent applications and subsequent changes of performance: evidence from time-series cross-section analyses on the firm level. *Research Policy*, 30(1), 143-157.
Gambardella, A., Harhoff, D., & Verspagen, B. (2011). The value of European patents. *European Management Review*, 5(2), 69-84.
Gayle, P. G. (2001). Market concentration and innovation: new empirical evidence on the schumpeterian hypothesis.
Haeussler, C., Harhoff, D., & Mueller, E. (2014). How patenting informs investors – the case of biotechnology. *Research Policy*, 43(8), 1286-1298.
Hoenen, S., Kolympiris, C., Schoenmakers, W., & Kalaitzandonakes, N. (2014). The diminishing signaling value of patents between early rounds of venture capital financing. *Research Policy*, 43(6), 956-989.
Hoenig, D., & Henkel, J. (2015). Quality signals? the role of patents, alliances, and team experience in venture capital financing. *Research Policy*, 44(5), 1049-1064.
Hu, M. C. (2008). Knowledge flows and innovation capability: the patenting trajectory of taiwan's thin film transistor-liquid crystal display industry. *Technological Forecasting & Social Change*, 75(9), 1423-1438.
Lei, X. P., Zhao, Z. Y., Zhang, X., Chen, D. Z., Huang, M. H., & Zheng, J., et al. (2013). Technological collaboration patterns in solar cell industry based on patent inventors and assignees analysis. *Scientometrics*, 96(2), 427-441.
Lemley, M. A. (2001). Rational ignorance at the patent office. Berkeley Olin Program in Law & Economics Working Paper, 95(1021), 1495-1532.
Reitzig, M. (2004). Improving patent valuations for management purposes—validating new indicators by analyzing application rationales. *Research Policy*, 33(6), 939-957.
Sapsalis, E., & Ran, N. (2006). Academic versus industry patenting: an in-depth analysis of what determines patent value. *Research Policy*, 35(10), 1631-1645.
Scherer, F. M., & Harhoff, D. (2004). Technology policy for a world of skew-distributed outcomes. *Research Policy*, 29(4), 559-566.
Smith, J. A. , & Cordina, R. . (2015). Patenting and the early-stage high-technology investor: evidence from the field. *R&D Management*, 45(5), 589-605.
Useche, D. (2014). Are patents signals for the IPO market? an EU–US comparison for the software industry. *Research Policy*, 43(8), 1299-1311.
Veer, T., & Jell, F. (2012). Contributing to markets for technology? a comparison of patent filing motives of individual inventors, small companies and universities. *Technovation*, 32(9-10), 513-522.
Zacharakis, A. L., & Meyer, G. D. (2005). The potential of actuarial decision models : can they improve the venture capital investment decision?. *Journal of Business Venturing*, 15(4), 323-346.
Zhang, Y., Qian, Y., Huang, Y., Guo, Y., Zhang, G., & Lu, J. (2017). An entropy-based indicator system for measuring the potential of patents in technological innovation: rejecting moderation. *Scientometrics*, 111(3), 1-22.
Zhang, Y., Shang, L., Huang, L., Porter, A. L., Zhang, G., & Lu, J., et al. (2016). A hybrid similarity measure method for patent portfolio analysis. *Journal of Informetrics*, 10(4), 1108-1130.
Zhou, H., Sandner, P. G., Martinelli, S. L., & Block, J. H. (2016). Patents, trademarks, and their complementarity in venture capital funding. *Technovation*, 47, 14-22.

Are younger researchers more internationally oriented than their senior colleagues?

A study of international research collaboration using co-authorship data

Kristoffer Rørstad, Dag W. Aksnes and Fredrik Piro

kristoffer.rorstad@nifu.no; dag.w.aksnes@nifu.no; Fredrik.piro@nifu.no
Nordic Institute for Studies in Innovation, Research and Education, Økernveien 9, 0653 Oslo, Norway

Abstract

This paper addresses the relationship between age and international research collaboration. The main research question is: do younger researchers collaborate more internationally than their senior colleagues? A common assumption is that the younger generations are more globally oriented in general than the older generations are. At the same time, senior researchers might have a much larger international network than their younger colleagues have. The study is based on a dataset consisting of 5,554 Norwegian researchers and their publication output during a three-year period (43,641 publications). The main indicator used to assess the question is the proportion of publications with international co-authorship. The study shows that the research field is by far the most important factor influencing the propensity to collaborate internationally. There are distinct generational differences where the older researchers have a lower intensity of international research collaboration than younger generations.

Introduction

To what extent does the age of scientists influence on their scientific performance? This question has been a recurring issue in the bibliometric literature. One major topic has concerned the relationship between publication productivity and age. Although the results of previous studies have not always been entirely consistent, it seems to be quite firmly established that there is a curvilinear relationship between age and productivity. The average production of publication increases with age and reaches a peak at some point during the career and then declines (see e.g. (Barjak, 2006; Gingras, Lariviere, Macaluso, & Robitaille, 2008; Gonzalez-Brambila & Veloso, 2007). The pattern has been found across many fields and nations.

A related issue concerns age and international collaboration. Are older researchers more inclined than younger to collaborate internationally or is there an inverse relationship? This issue has received much less attention, and the main contributions stem from questionnaire surveys rather than bibliometric data. In a previous Norwegian study, academic staff were asked to report if they had collaborated with foreign colleges during a three-year period (Kyvik & Olsen, 2008). This study showed that the propensity to collaborate internationally were lowest for the youngest and oldest age-groups (below 35 years old and above 60) and similar findings were reported in an updated survey (Kyvik & Aksnes, 2015; Kyvik & Reymert, 2017).

In order to provide more knowledge on this issue, we address the issue bibliometrically in this study. This will enable us to analyze not only the proportion of researchers involved in such collaboration by age, but also the intensity of their international orientation. Here, international co-authorship (publications with authors affiliated with institutions in more than one country) is applied as a proxy or indicator of international collaboration. Generally, bibliometric data is a useful and valuable data source for studying international collaboration, although there are also limitations attached to using co-authorship as measure of collaboration (Katz & Martin, 1997).

It is well known that the extent of international research collaboration has increased significantly during the recent decades (Adams, 2012; Wagner, Whetsell, & Leydesdorff, 2017). In many countries, the majority of scientific publications are now internationally co-authored. In Norway, this proportion has grown from approximately 20 per cent during the 1980s to more than 60 per cent recent years (Web of Science indexed publications). At the same time, there are large differences in the collaboration rates across countries and fields (National Science Board, 2018). For example, the proportion of publications involving international co-authorship is much higher in the natural sciences than in the social sciences and humanities.

In this study we apply a generational perspective to the issue of international collaboration. Generally, the behaviour of individuals is influenced by the generation they belong to (Kyvik & Aksnes, 2015). This means that younger and older scientists may deviate in their research practice due to differences in the socio-cultural influences at different times. Generational differences may be observed because researchers have been trained and socialized into the academic profession in divergent ways. At the same time, younger and older researchers typically have different roles in the research system, and varying experience and competence which might influence on their collaboration patterns. Against this background, based on a dataset of Norwegian researchers, we will assess the following hypotheses and possible explanations:

a) Senior researchers are more internationally oriented than their younger colleagues are. During their career they have been able to establish stronger and more extensive international networks and the older, experienced researchers would also be more attractive as collaborative partners. Thus, one would expect that this is reflected in the collaboration patterns.
b) Senior researchers are less internationally oriented than their younger colleagues are. The older researchers were trained at a time when international collaboration was much less common than today. This attitude may to some extent still characterise and influence their collaboration practice. As a result, one may find generational differences in the international collaboration patterns.

While most bibliometric studies are based on the Web of Science and Scopus databases, we use a national database which has a complete coverage of the publication output. By this we are able to provide a much better representation of the social sciences and the humanities in particular, which are less well covered in common databases, e.g. Scopus and Web of Science.

Methods and data

The study is based on the bibliographic *Cristin* database (the Norwegian Science Index) that has been developed as part of a current research information system for all public research institutions in Norway. The database covers all peer-reviewed scientific and scholarly publication output, including books, edited volumes and conference series (Piro, Aksnes, & Rorstad, 2013). The system secures complete, verifiable and structured data for bibliometric analysis. Of particular importance for the study of scientific collaboration is that all authors and addresses are indexed, including country as a controlled term. In this study, we have analyzed publications from the three-year period 2015-2017.

As a source of information on individual characteristics of the persons (gender, age, position, and institution), the data in the bibliographic database was coupled with another database, the *Norwegian Research Personnel Register*. This database contains individual data for all researchers in the Higher Education Sector and Institute Sector in Norway. The individuals were classified in 10-year age cohorts defined as the aggregate of individuals who were born in the same time interval.

The data material consists of 5,554 researchers from the four largest universities in Norway (University of Oslo, University of Bergen, The Arctic University of Norway and The Norwegian University of Science and Technology). The study is limited to professors, associate professors, postdocs and PhD students with at least one publication during the time period analyzed. The researchers were assigned to five broad fields (Social sciences, Humanities, Natural sciences, Engineering and technology and Medical and health sciences), based on the field distribution of their publication output. Their publication output during the period 2015-2017, in total accounts for 43,641 publications.

As a measure of international collaboration, two indicators are applied. First, the share of researchers who have been involved in international collaboration measured by co-authorship (i.e. have published at least one publication involving international co-authorship). This is a measure of whether the individuals have collaborated internationally or not and corresponds to the indicator used in the study by Kyvik and Olsen (2008). Second, the proportion of the publications involving international co-authorship. This is a measure of the intensity of the international collaboration (Abramo, D'Angelo, & Murgia, 2013). The unit for the analyses is the individual researchers. In other words, all individuals count equally as one unit in the analysis regardless of how many publications they have published. By this, we avoid that the analysis is biased towards highly productive researchers.

The analyses are carried out at the level of major scientific fields. The academic position of the researchers is used as another independent variable (due to the different publications patterns across these variables).

Results

Overall, 62 per cent of the researchers were involved in international collaboration measured by co-authorship. In other words, almost two thirds have published at least one publication with international co-authorship during the three-year period. However, there are large variations in the proportions across fields and academic positions.

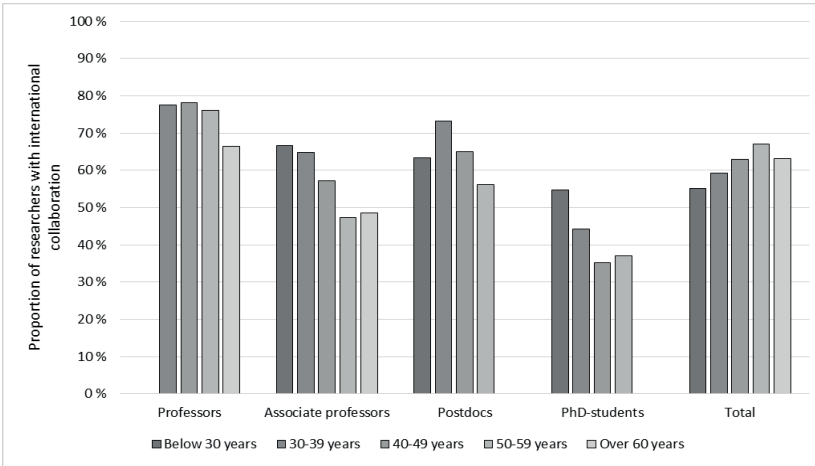


Figure 1 Proportion researchers involved in international collaboration by academic positions and age groups (N=5,554)*

*) Figures are not shown for categories with less than ten researchers.

Figure 1 shows the proportions using scientific position and age group as independent variables. For the professors, associate professors and PhD-students, the propensity to

collaborate internationally is declining with age, albeit with some variations in the age patterns. For instance, 78 per cent of the professors in the age group 30-39 years were involved in international collaboration, while corresponding figure for the professors above 60 years was 66 per cent

Interestingly, the total shows an opposite age pattern where an increasing proportion of the researchers have been involved in international collaboration. However, this can be explained by the different composition of researchers across academic positions. PhD students, which have the lowest proportions of international collaboration, constitute the large majority of the individuals below 30 years, while professors are over-represented in the older age groups.

In the remaining part, we will focus on the alternative indicator measuring the intensity of the international collaboration. Figure 2, which shows the average proportion of international co-authorship per individual by scientific position and age group, basically provides a similar picture as Figure 1. In general, younger researchers have higher degrees of international co-authorship than their older colleagues.

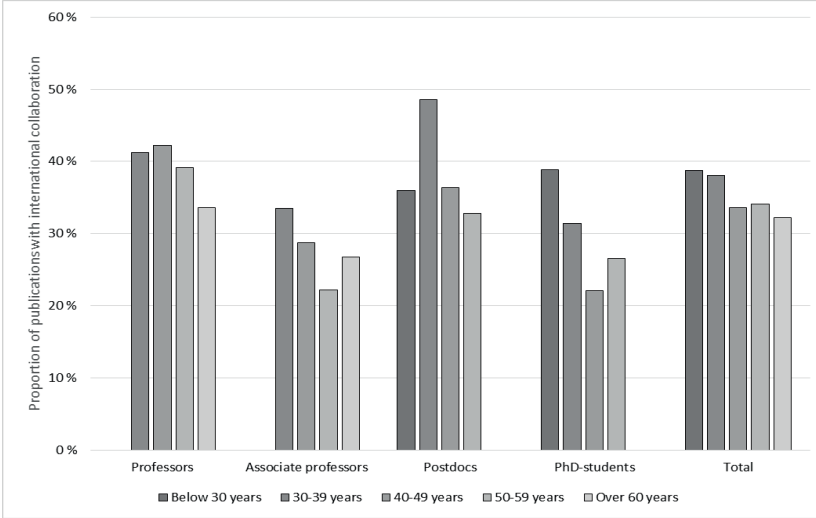


Figure 2 Average proportion of international co-authorship per individual by academic position and age group (N=5,554)*

*) Figures are not shown for categories with less than ten researchers.

Analysed by major field and age group, we find a distinct declining age pattern for the professors in the Humanities, the Natural sciences and Engineering and technology, while the Social sciences and Medicine and health show a curvilinear age pattern (Figure 3). The figure also shows that international collaboration is much more frequent in the Natural sciences, Medical and health sciences and Technology compared with Humanities and Social sciences. This holds for all age groups.

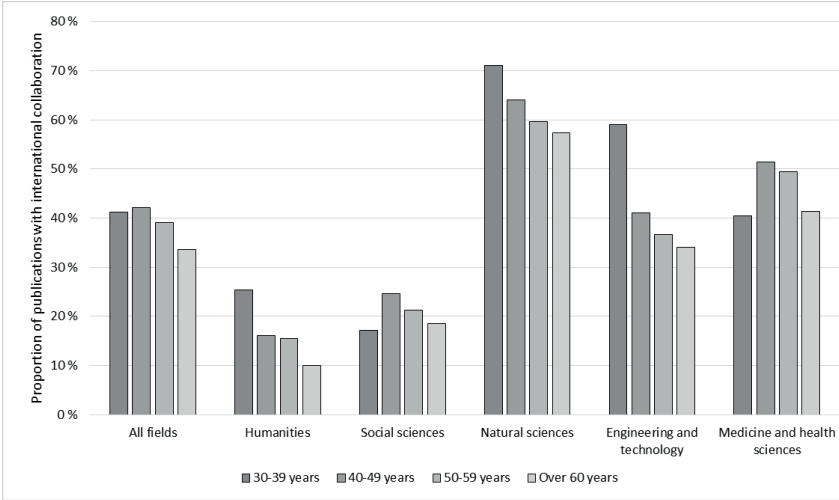


Figure 3 Average proportion of international co-authorship for per individual for professors by fields and age group (N=2,068)*

*) Figures are not shown for categories with less than ten researchers.

A declining pattern of international collaboration rate by age is also found when using individual years rather than age cohorts as unit of analysis. This is shown in the scatterplot in Figure 4, where the average degree of international co-authorship is plotted for each year group, i.e. one dot per age year. The R-squared value is 0.31 indicating a weak to moderate negative linear relationship.

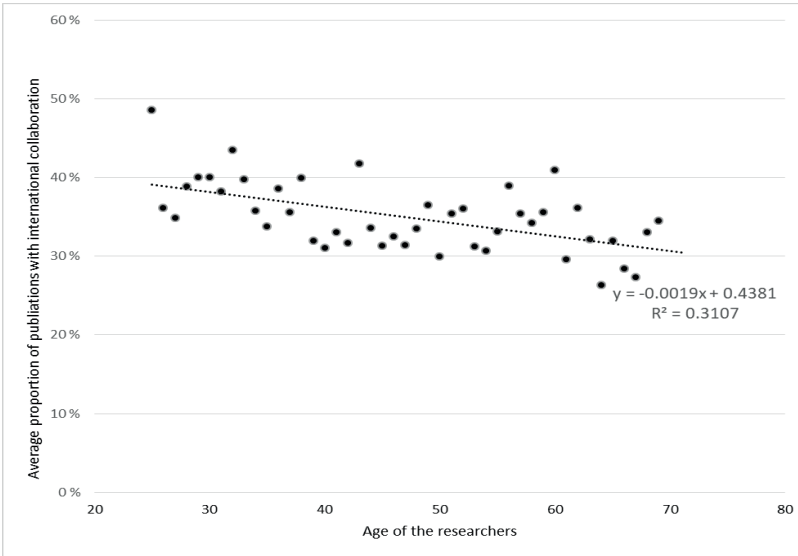


Figure 4 Scatter plot of average proportion of international co-authorship for per individual by age (N=5,554)

Discussion & conclusions

Our preliminary analyses on an aggregated level have shown that there are distinct age differences in international research collaboration where the oldest researchers tend to be less involved in such collaboration than the younger generations. This holds across both indicators applied to assess the question: the proportion of researchers involved in international collaboration and the intensity of their international collaboration. Although these results are country specific for Norway, we assume that the findings may be generalizable.

In the introduction we outlined some possible explanations for generational differences in this respect. We suggest that the differences may be explained by younger researchers being more cosmopolitan in their research practice than the older staff. This again might be linked to differences in the socio-cultural factors influencing the researchers when being trained into the academic profession. The older researchers were trained at a time when international collaboration was much less common than today, which still to some extent influence their collaboration practice.

This is a research in progress paper. In the proceeding work, we will analyse additional dimensions such as how differences in publication productivity at an individual level affect the collaboration patterns and the role of collaboration more generally measured as average number of authors per paper. Since age is correlated with academic position and to some extent gender, we will include these variables in our proceeding work to be able to isolate the age effect in this research question.

References

Abramo, G., D'Angelo, C. A., & Murgia, G. (2013). Gender differences in research collaboration. *Journal of Informetrics, 7*(4), 811-822. doi:10.1016/j.joi.2013.07.002

Adams, J. (2012). The rise of research networks. *Nature, 490*(7420), 335-336. doi:10.1038/490335a

Barjak, F. (2006). Research productivity in the internet era. *Scientometrics, 68*(3), 343-360. doi:10.1007/s11192-006-0116-y

Gingras, Y., Lariviere, V., Macaluso, B., & Robitaille, J. P. (2008). The Effects of Aging on Researchers' Publication and Citation Patterns. *Plos One, 3*(12). doi:10.1371/journal.pone.0004048

Gonzalez-Brambila, C., & Veloso, F. M. (2007). The determinants of research output and impact: A study of Mexican researchers. *Research Policy, 36*(7), 1035-1051. doi:10.1016/j.respol.2007.03.005

Katz, S. J., & Martin, B. R. (1997). What is research collaboration? *Research Policy, 26*(1), 1-18.

Kyvik, S., & Aksnes, D. W. (2015). Explaining the increase in publication productivity among academic staff: a generational perspective. *Studies in Higher Education, 40*(8), 1438-1453. doi:10.1080/03075079.2015.1060711

Kyvik, S., & Olsen, T. B. (2008). Does the aging of tenured academic staff affect the research performance of universities? *Scientometrics, 76*(3), 439-455. doi:10.1007/s11192-007-1767-z

Kyvik, S., & Reymert, I. (2017). Research collaboration in groups and networks: differences across academic fields. *Scientometrics, 113*(2), 951-967. doi:10.1007/s11192-017-2497-5

National Science Board. (2018). *Science and Engineering Indicators 2018*. Retrieved from Alexandria, VA: https://www.nsf.gov/statistics/indicators/

Piro, F. N., Aksnes, D. W., & Rorstad, K. (2013). A macro analysis of productivity differences across fields: Challenges in the measurement of scientific publishing. *Journal of the American Society for Information Science and Technology, 64*(2), 307-320. doi:10.1002/asi.22746

Wagner, C. S., Whetsell, T. A., & Leydesdorff, L. (2017). Growth of international collaboration in science: revisiting six specialties. *Scientometrics, 110*(3), 1633-1652. doi:10.1007/s11192-016-2230-9

Does the Gini coefficient of a journal's citations increase over time?

Ronald Rousseau[1,2], Xiaojun Hu[3], Huiying Du[4], Yujie Peng[4], Lin Zhang[5]

[1] ronald.rousseau@kuleuven.be
KU Leuven, MSI, ECOOM, B-3000 Leuven (Belgium)

[2] *ronald.rousseau@uantwerpen.be*
University of Antwerp (UA), Faculty of Social Sciences, B-2020 Antwerp, (Belgium)

[3] *xjhu@zju.edu.cn*
Medical Information Centre, and Department of Neurology of Affiliated Hospital 2, Zhejiang University School of Medicine, Hangzhou 310058 (China)

[4] *dhy9596@126.com, pyj8176@126.com*
North China University of Water Resources and Electric Power, Zhengzhou, China

[5] *zhanglin_1117@126.com*
School of Information Management, Wuhan University, China

Abstract

In this contribution the authors investigate the evolution of inequality among received citations for articles published in the same journal. We study the slope – increasing or decreasing – of the Gini coefficient of a journal's citations over a period of 30 years. As a result we find that for most journals this slope is increasing, pointing to an increased inequality among citations, but this increase itself is not necessarily decreasing. The journals *Nature* and *Science* show, surprisingly, a different behaviour. As a limitation we point out that we only investigated one publication year (1985) and selected journals belonging to just a few fields.

Introduction

Theoretical considerations

Consider a set of related articles of the same age. We want to study the evolution of inequality in received citations of these articles. As a base model we assume that at any given time further citations occur in proportion to the number of already received citations. This assumption implies that the ranking of these articles is unchanged. In that case the Lorenz curve of received citations and the corresponding Gini index also stay invariant. Yet, because we know that in bibliometrics most phenomena, and in particular citations, follow a success-breeds-success pattern (Price, 1976), or stated otherwise: the most-cited articles enjoy a Matthew effect (Merton, 1968, 1988; Mahbuba & Rousseau, 2011), this base model will not occur in reality. One expects that the most cited articles are proportionately more cited in future publications than the less cited ones. Also then the ranking of these articles stays unchanged. This leads to a Lorenz curve that becomes situated further away from the diagonal and hence an increase in the value of the Gini index over time. Stated otherwise: citation curves are right skewed. Finally, as the yearly number of citations received by a journal and by its articles generally decreases over time, the increase in the Gini index is not linear but concave. Consequently we claim that the ideal, theoretical evolution of the citation Gini index of a given set of articles looks like shown in Figure 1. The horizontal direction refers to time while the vertical direction refers to the value of the Gini index. As we do not make a claim on the values of the Gini index we do not show a vertical axis.

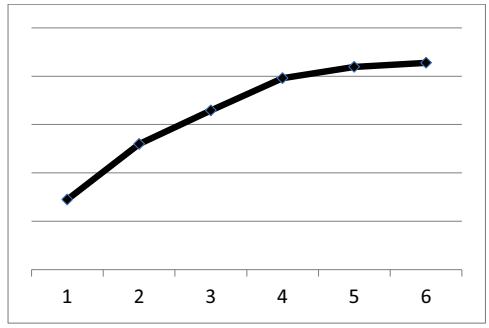


Figure 1. Ideal evolution of the Gini index.

A more realistic approach

Now we come to some more realistic considerations. The theory as exposed above implies that the ranking of articles, according to received citations does not change. Yet, received citations, being the consequence of human behaviour, do not grow according to an invariable law. In particular if this ranking is highly dynamic due to e.g., delayed recognition, flashes-in-the-pan, under-cited influential articles, retractions and extremely highly-cited articles (van Raan, 2004; Fanelli, 2014; Hu & Rousseau, 2016), the inequality in received citations may even decrease. Hence we expect that if such special cases have a minor influence then a journal's citation Gini index is increasing over time, but not necessarily according to a fluent concave curve as suggested in Fig. 1. Moreover, if some irregularities happen we expect that a best-fitting linear regression line of Gini index values still has a strictly positive slope. This leads to the following hypotheses:

H1: Inequality in received citations of journals increases over time

H2. Inequality in received citations of journals increases in a concave way

where inequality is measured through the Gini index. Only if a journal meets the first hypothesis we investigate further to find out if also the second one is met. As there may be many confounding factors we will only include journals with a (relatively) large number of publications restricted to documents of article-type, according to the Web of Science (WoS).

Methods and data

We consider journals included in the Web of Science (WoS) published in the year 1985. For a given journal we collect for each article the number of citations received over the periods [1985-1990]; [1985-1995]; [1985-2000]; [1985-2005]; [1985-2010]; [1985-2015]. For each of these periods the Gini index for the inequality in received citations is calculated. This leads to 6 values (6 data points) for each journal. Then it is checked if the best fitting regression line has a positive slope. As we have only six data points, it may easily happen that there is an outlier, e.g. a sudden unexplained increase or decrease in the value of the Gini index. For this reason we do not calculate the slope of a best fitting line in the ordinary least squares (OLS) sense, but of a non-parametric best-fitting line as described in (Helsel & Hirsch, 2002; Rousseau et al., 2018). If it can be accepted that the slope is positive, according to a Kendall's S test (equivalently a Kendall tau test) with $p = 0.1$, we continue and study the differences in Gini indices (5 data points) and again calculate the non-parametric slope to see if it is negative, where we use the same statistical test. As we only have a small number of data points we use Table B8 in (Helsel & Hirsch, 2002) to decide on statistical significance.

We recall that the formula to calculate a Gini index of an array $X = (x_j)_{j=1,\ldots,N}$ of length N, denoted as g(X), is (Rousseau et al., 2018):

$$g(X) = \frac{1}{N}\left(N + 1 - \frac{2}{TOT}\sum_{j=1}^{N} j x_j \right)$$

In this formula the array X is ranked in decreasing order and TOT denotes the sum of all its values: $TOT = \sum_{i=1}^{N} x_j$.

Our investigation has an explorative nature. It will become clear that we do not strive for (and will not obtain) clearly delineated results.

Results

In this "Work in Progress" we only report on some journals in the categories Physics (12 journals), Neurosciences (6 journals), Clinical Medicine (7 journals) plus *Nature* and *Science*. We say that H1 is satisfied if a journal's Gini coefficients show a strictly positive slope over time. Similarly, we say that H2 is satisfied if the differences in the consecutive values of a journal's Gini indices have a strictly negative slope. Results are shown in Table 1.

Table 1. Results.

Journal	Category	Slope Gini values positive	Slope differences in Gini values are negative
AMERICAN JOURNAL OF PHYSICS	Physics	YES	NO
JETP LETTERS	Physics	NO	
JOURNAL OF CRYSTAL GROWTH	Physics	NO	
JOURNAL OF MAGNETISM AND MAGNETIC MATERIALS	Physics	YES	NO
JOURNAL OF PHYSICS A – MATHEMATICAL AND GENERAL	Physics	YES	YES
JOURNAL OF THE PHYSICAL SOCIETY OF JAPAN	Physics	YES	NO
LETTERE AL NUOVO CIMENTO	Physics	YES	NO
PHYSICA STATUS SOLIDA A – APPLIED RESEARCH	Physics	NO	
PHYSICAL REVIEW LETTERS	Physics	YES	NO
PHYSICS LETTERS A	Physics	YES	YES
PROGRESS OF THEORETICAL PHYSICS	Physics	YES	NO
THIN SOLID FILMS	Physics	YES	YES
BEHAVIOUR & BRAIN SCIENCES	Neurosciences	YES	YES
BRAIN	Neurosciences	YES	YES
JOURNAL OF CEREBRAL BLOOD FLOW AND METABOLISM	Neurosciences	YES	YES
JOURNAL OF NEUROCHEMISTRY	Neurosciences	YES	YES
JOURNAL OF NEUROSCIENCE	Neurosciences	YES	NO
NEUROPHARMACOLOGY	Neurosciences	YES	YES
ANNALS OF INTERNAL MEDICINE	Clinical Medicine	YES	NO
BRITISCH MEDICAL BULLETIN	Clinical Medicine	YES	NO
JAMA	Clinical medicine	NO	

JOURNAL OF LABORATORY AND CLINICAL MEDICINE	Clinical Medicine	YES	YES
LANCET	Clinical Medicine	NO	
MAYO CLINIC PROCEEDINGS	Clinical Medicine	NO	
NEW ENGLAND JOURNAL OF MEDICINE	Clinical Medicine	YES	NO
NATURE	Multidisciplinary	NO	
SCIENCE	Multidisciplinary	YES	NO

Among these 27 journals 9 (one third) have the expected shape, hence satisfy hypotheses 1 and 2: Gini values are increasing over time and differences decrease over time. Another eleven have increasing Gini values (satisfy the first hypothesis), but differences do not decrease over time; while seven do not even have increasing Gini values (do not satisfy any of our hypotheses). Journals in the field of Neurosciences follow the expected form the best, while those in Internal Medicine deviate most from our expectations. Figures 2, 3 and 4 show the evolution of Gini coefficients for the journals *Thin Solid Films*, *Science* and *Nature*, illustrating the three cases. Numbers 1 to 6 refer to the six periods under study.

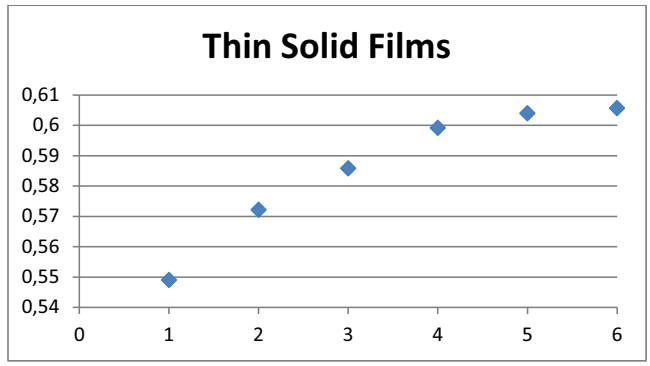


Figure 2. The evolution of Gini indices for the journal *Thin Solid Films* is as expected.

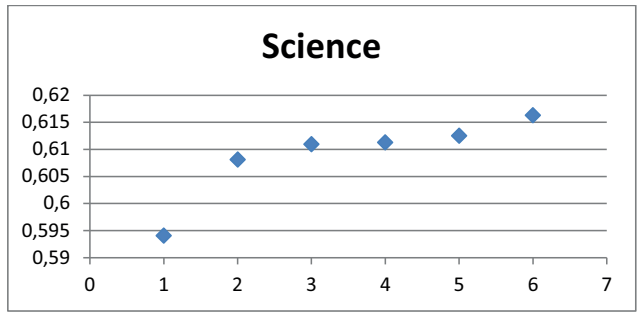


Figure 3. Gini indices for *Science* are increasing but not in a concave way.

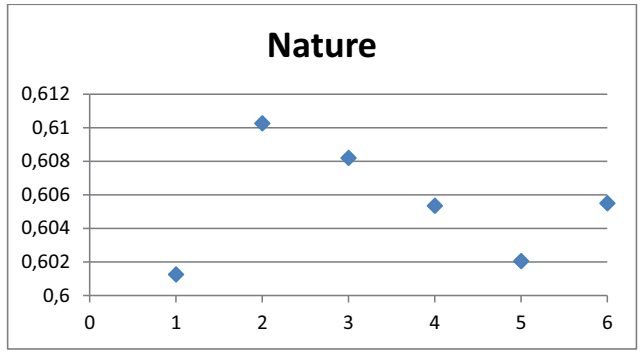


Figure 4. Gini indices for *Nature* are not increasing.

Conclusion and discussion

Our results lead to the conclusion that at least for the investigated journals the inequality of received citations over publications of article type is mostly, but certainly not always, increasing over time. The Gini values (over a thirty year period) we provided for these journals expand our knowledge about the skewness of citations received by articles published in the same journals (Rousseau, 2014). They provide concrete values.

This work in progress has obvious limitations: we only investigated one publication year (1985) and a small number of selected journals belonging to a few WoS categories. Even then our results are not clear-cut but just point at a general tendency. Surprisingly, the journals *Nature* and *Science,* although generally considered as *similar,* show a totally different behaviour in terms of inequality of received citations among their articles.

This work is based on a presentation during the COLLNET conference at the University of Kent, Canterbury, 2017.

Acknowledgments

This work was supported by the National Natural Science Foundation of China: Grants 71573225 and 71573085; and the Excellence Scholarship in Social Science in HeNan Province (No.2018-YXXZ-10). The authors thank Xian Li and Xiaoyue Hu from Zhejiang University for help in data collection.

References

Fanelli, D. (2014). Publishing: rise in retractions is a signal of integrity. *Nature*, 509(7498), p.33.

Helsel, D.R. & Hirsch, R. M. (2002). Statistical methods in water resources. Chapter A3 in *Hydrologic Analysis and Interpretation*. U.S. Department of the Interior.

Hu, XJ. & Rousseau, R. (2016). Scientific influence is not always visible: The phenomenon of under-cited influential publications. *Journal of Informetrics*, 10(4), 1079-1091.

Mahbuba, D. & Rousseau, R. (2011). The Matthew effect and a relation with concept symbols and defaults. *Annals of Library and Information Studies,* 58(4), 335-345.

Merton, R.K. (1968). The Matthew effect in science. *Science*, 159(3810), 56-63.

Merton, R.K. (1988). The Matthew effect in science II. Cumulative advantage and the symbolism of intellectual property. *Isis*, 79(4), 606-623.

Price, D. J. de Solla (1976). A general theory of bibliometric and other cumulative advantage processes. *Journal of the American Society for Information Science*, 27(5), 292-306.

Rousseau, R. (2014). Skewness for journal citation curves. In: (E. Noyons, Ed.) *Proceedings of the Science and Technology Indicators Conference 2014 Leiden Context Counts: Pathways to Master Big and Little Data*. Leiden: Universiteit Leiden, 498-501.

Rousseau, R., Egghe, L. & Guns, R. (2018). *Becoming metric-wise. A bibliometric guide for researchers.* Kidlington: Chandos (Elsevier).
Van Raan, A.F.J. (2004). Sleeping beauties in science. *Scientometrics*, 59(3), 467-472.

A New Algorithm for Zero-Modified Models Applied to Citation Counts

Marzieh Shahmandi[1], Paul Wilson[2], and Mike Thelwall[3]

[1] *m.shahmandihounejani@wlv.ac.uk*

[2] *pauljwilson@wlv.ac.uk*

[3]m.thelwall@wlv.ac.uk

Statistical Cybermetrics Research Group, University of Wolverhampton (United Kingdom)

Abstract

Finding statistical models for citation count data is important for those seeking to understand the citing process or when using regression to identify factors that associate with citation rates. As sets of citation counts often include more or less zeros (uncited articles) than would be expected under the base distribution, it is essential to deal appropriately with them. This article proposes a new algorithm to fit zero-modified versions of discretised log-normal, hooked power-law and Weibull models to citation count data from 23 different Scopus categories from 2012. The new algorithm allows the standard errors of all parameter estimates to be calculated, and hence also confidence intervals and p-values. This algorithm can also estimate negative zero-modification parameters corresponding to zero-deflation (fewer uncited articles than expected). The results find no universal best model for the 23 categories and a given dataset may be zero-inflated relative to one model, but zero-deflated relative to another.

Introduction

It is important to identify models that fit citation distributions well for several reasons. A correct model can be used to identify anomalous sets of articles that are not fitted well by a model based upon an incorrect distribution and can help with the design of effective impact indicators. It is important when performing regression analyses to identify factors that influence citations. Also, confidence intervals for, say, model coefficients, based upon a poorly fitting model may be either too wide or too narrow, leading to incorrect estimates of possible effect sizes. It sometimes happens that the number of 0's in a given dataset are not fitted well by a distribution. This problem can often be remedied by fitting a zero-inflated or a zero-deflated (i.e. a zero-modified) distribution that allows the predicted number of zeros to approximately equal the number of zeros in a dataset.

A previous study fitted zero-inflated versions of the discretised log-normal and hooked power law distributions to citation count data from 23 Scopus categories, finding that zero-inflation occurred in the vast majority of cases (Thelwall, 2016). The zero-inflation was hypothesised to be a consequence of "inherently unciteable articles", such as magazine articles. Zero-counts due to unciteability are an example of "perfect" or "structural" zeros: data that are constrained to be zeros due to some feature of the data generating process. In contrast, other zeros are referred to as non-perfect or count zeros. In this context a non-perfect zero would be a paper that is citeable, but has not been cited. In essence, zero-inflated models seek to estimate the proportion of perfect zeros present in data, and fit a count distribution to the remaining data.

A less well-studied phenomenon is zero-deflation, where data is well-fitted by a given count distribution, but there are less zeros present in the data than would be expected under the distribution. Zero-deflation may arise for citation counts from the Web of Science (WoS), Scopus or any other citation database with selective inclusion criteria because uncited articles may be less likely to be indexed. WoS and Scopus have poorer coverage of non-English journals than of English journals so non-English journals may contribute to zero-deflation. This may be

particularly relevant for fields containing nation-specific agricultural, legal, culture, or politics research.

Whilst previous studies have fitted zero inflated distributions to citation count data, none have fitted zero-deflated or zero-modified distributions to citation count data. This paper introduces zero-modified versions of the hooked power law and discretized log-normal distributions previously shown to fit citation data well (Thelwall, 2016), and also zero-modified versions of the discrete Weibull distribution. Brzezinski (2014) discusses the use of the discrete Weibull distribution to model citation counts, the discrete Weibull distribution is capable of modelling highly skewed count data with more zeros and thus is a good candidate model for citation count. Discrete Weibull distributions may be fitted to data using the R-Package DWreg (Vinciotti, 2016). The pure power law distribution is not considered because it usually requires low cited articles to be ignored for fitting and therefore is not a credible citation distribution. This paper also introduces an algorithm that fits both negative and positive zero-modification parameters, and determines the standard errors of the zero-modification (and other) parameters, which in turn enables the calculation of confidence intervals for these parameters, and the performance of statistical tests on them. The algorithms are tested on a sample set of citation data from 23 fields to assess the extent to which the new distributions fit citation count data.

Distributions

Hooked Power Law

The hooked power law (Thelwall, 2016) is a generalised version of the power law model (Pennock, David & Flake et al., 2002). The hooked power law has a probability mass function:

$$f(x;B,\alpha) = \begin{cases} A(B+x)^{-\alpha} & x = 0,1,2,\ldots \\ 0 & otherwise \end{cases}$$

where B and α are model parameters, and A is a constant chosen so that $\sum_{x=0}^{\infty} f(x;B,\alpha) = 1$.

Discretized Log-normal

A (continuous) random variable is log-normally distributed if its logarithm is normally distributed. It has probability density function:

$$f(x;\mu,\sigma) = \frac{1}{x\,\sigma\sqrt{2\pi}} \exp\left(-\frac{(\ln(x)-\mu)^2}{2\sigma^2}\right), \qquad x > 0, \sigma > 0, \mu \in (-\infty, +\infty). \tag{1}$$

To discretise the distribution, (i.e., convert it into a form that models the situation where x is a positive integer), integrate $f(x;\mu,\sigma)$ over unit intervals about positive integer values of x, and divide by $K = \int_{0.5}^{\infty} f(x;\mu,\sigma)dx$, where f is as at (1) above. Thus, the probability mass function of the discretised log-normal distribution is:

$$g(x;\mu,\sigma) = \begin{cases} \frac{1}{K}\int_{x-0.5}^{x+0.5} f(x;\mu,\sigma)dx & x = 0,1,2,3,\ldots \\ 0 & otherwise \end{cases}$$

Discrete Weibull

The discrete Weibull distribution has probability mass function:

$$f(x;\mu,\sigma) = \begin{cases} q^{x^\beta} - q^{(x+1)^\beta} & x = 0,1,2,3,\ldots \\ 0 & otherwise \end{cases}$$

where $0 < q < 1$ and $\beta > 0$.

Zero-modified models

A zero-modified model (see, for example, Dietz and Böhning, 2000) has the probability mass function:

$$f(x;\Theta) = \begin{cases} \omega + (1-\omega)f^*(x;\Theta) & x = 0 \\ (1-\omega)f^*(x;\Theta) & x = 1,2,3,\ldots \\ 0 & otherwise \end{cases} \qquad (2)$$

Where Θ is a set of parameters and $f^*(x;\Theta)$ is a probability mass function. for negative ω the distribution is known as a zero-deflated distribution and for positive ω, it is known as a zero-inflated distribution. For $\omega = 0$ the model reduces to the non-modified model, f^*, and if $\omega = 1$ then the data would consist entirely of zeros. The zero-inflated model can be considered as a method of modelling the number of excess zeros (zero counts greater than expected under the model f^*), which can stem from two distinct processes, one process where zeros occur by chance, in the same manner as 1s, 2s, . . . occur; and another process by which some data are constrained to be zeros (perfect or structural zeros).

For a zero-deflated model, $\omega < 0$, but may take values < -1. To see note that

$$\begin{aligned} f(0;\omega,\Theta) \geq 0 &\Leftrightarrow \omega + (1-\omega)f^*(0;\Theta) \geq 0 \\ &\Rightarrow \omega\big(1 - f^*(0;\Theta)\big) + f^*(0;\Theta) \geq 0 \\ &\Rightarrow \omega > \frac{f^*(0;\Theta)}{1 - f^*(0;\Theta)} \end{aligned}$$

For example, if f^* is a Poisson distribution with parameter 0.5 then $f^*(0;0.5) = \exp(-0.5) = 0.6065$ and hence ω is valid provided

$$\omega \geq -\frac{0.6065}{1 - 0.6065} = -1.54$$

The interpretation of negative values of ω is not as straightforward as those of positive values. The most straight forward interpretation is to regard $1 - \omega$ as the proportionate increase in the expected number of observed positive values. For example, if $\omega = -1.5$, then we would expect to observe approximately $1 - (-1.5) = 2.5$ times more 1's, 2's, 3's etc in the data than we would in the non-modified model. Zero-deflation in data is usually as a consequence of some zero-counts not being included in the data. For example, Dietz and Böhning (2000) modelled zero-deflated DMFT index data from a dental epidemiological study previously published by Mendonca (1995). Specifically, the DMFT index quantifies the dental status of an individual through a count of "Decayed, Missing and Filled Teeth", and it was noted that an "incorrect sampling procedure" had led to the non-inclusion of some children whose score was zero.

Data and Methods

The data used in this article consist of citation counts for journal articles published in 2012 from 23 Scopus categories with up to 5000 journal articles for most of the categories. The citation counts to date were downloaded from Scopus in November 2017. The 5000 articles are the most recent 5000 for categories with more than 5000 articles. Whether a complete set or the most recent set of articles, this provides a coherent collection of articles with 5-6 years of citations.

A previously published algorithm fits zero-inflated discrete log-normal and zero-inflated hooked power law models to covariate free data (the zero-inflation parameter is estimated to two decimal places) (Thelwall, 2016). This model is easily extendable to zero-inflated versions of any count model, but is unable to fit negative zero-modification parameters. In this article, we include R-code for an algorithm that will enable the fitting of negative (and positive) values

of ω, it also will estimate the value of ω to many decimal places and is much faster. R code to fit the models discussed in this paper is available online[1].

This algorithm is based upon maximization of the log-likelihood of the relevant zero-modified models via the *optim* command of R. The optim function offers a number of different optimisation algorithms including conjugate gradient, quasi-Newton, Nelder-Mead and simulated annealing. The default method is a derivative-free Nelder-Mead algorithm that is a method for solving high-dimensional linear optimisation problems with constraints that is not sensitive to discontinuities in the likelihood surface.

The above mentioned algorithms have advantages over techniques such as Newton-Raphson and Fisher Scoring. In particular, they optimise log-likelihood function according to the parameters simultaneously as opposed to individually, such techniques have been around a long time, but have only become practical in recent years due to improved computing power.

The model with the greatest log-likelihood usually being considered as the most appropriate model among those being considered, log likelihood does not take into account the number of parameters being estimated in the given model however.

The *optim* command has the added advantage of returning an estimate of the matrix of second order partial derivatives of the log-likelihood function, $l(f)$ corresponding to the probability mass function, f. This matrix is known as the Hessian Matrix (Faraway, 2005) of $l(f)$, and is of importance as it may be shown that the diagonal entries of its inverse are proportionate to the standard errors of the parameter estimates. This is especially useful as it enables the calculation of confidence intervals for the zero-modification parameter (as well as any other parameters), values between the interval's limits are compatible with the data, given the statistical assumptions used to compute the interval; and the performance of hypothesis tests concerning the parameters, in particular it enables test of $H_0: \omega = 0$ to determine whether there is statistical evidence of zero-modification in the data. Whilst following the publication by the American Statistical Association (Wasserstein & Lazar, 2016) of guidelines concerning the misuse of p-values and confidence interval has led to considerable debate about the use of confidence intervals and p-values, the guidelines are primarily concerned with the misuse use of p-values and confidence intervals and far from advise there abandonment. Indeed, the guidelines specifically state that "P-values can indicate how incompatible the data are with a specified statistical model''.

Several tests exist to test for zero-modification, including likelihood ratio tests, score tests, and the Wilson-Einbeck test (Wilson & Einbeck, 2018). Note that whilst the Vuong test for non-nested models has been used as a test of zero-inflation, this is erroneous (Wilson, 2015). This paper uses the Wald test (Wasserman, 2006) to test: $H_0: \omega = 0$ against the alternative: $H_1: \omega \neq 0$ with $W = \frac{\widehat{\omega}}{Se(\omega)}$ where Se is the standard error of the maximum likelihood estimate of ω. We employ the Wald test as it directly tests the significance of the estimate of the zero-modification parameter without necessitating the fitting of the non-zero-modified model.

Finally, for assessment of the fitted model, Akaike information criterion (AIC) is used to show whether one model fits the data set better than another, when the models in question contain differing numbers of parameters or predictor variables (Akaike, 1974).

[1] The R source code is available at https://doi.org/10.6084/m9.figshare.7643093.v1

Results

Proportions of uncited articles

Uncited articles are far more common is some disciplines than in others (Figure 1). Cultural Studies, Economics & Econometrics, Health Social Science, and Pharmaceutical science have the greatest proportions of zero counts respectively. The radius of the circle is proportional to the number of articles with zero-citations for each discipline. It appears that in subjects such as Pharmaceutical Science large numbers of uncited articles arise from publications that might be regarded as magazines rather than journals being included in the database.

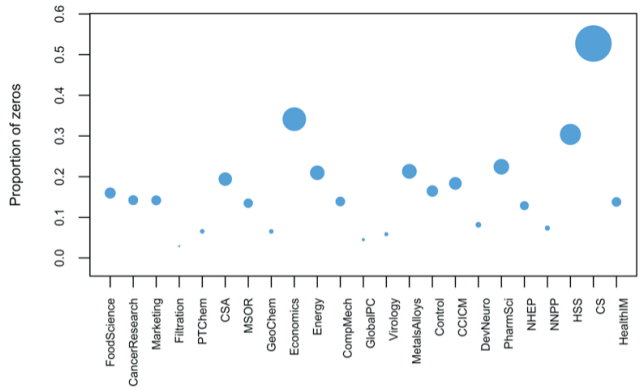


Figure 1. The proportions of uncited articles (zeros) in citation data from 23 Scopus categories.

Zero-modified discretised log-normal distribution

The zero-modification parameter estimates from the discretised log-normal distribution are all positive, the largest estimates being for Health Social Science and Economics and the smallest for Filtration & Separation and Global &Planetary Change (Figure 2, see also Table A1). The zero-inflation parameter estimates for 22 of the 23 subjects are significant at a level of significance of α=0.05, with only Health Information Management returning a non-significant estimate. There is almost universal zero-inflation relative to the discretized log-normal distribution.

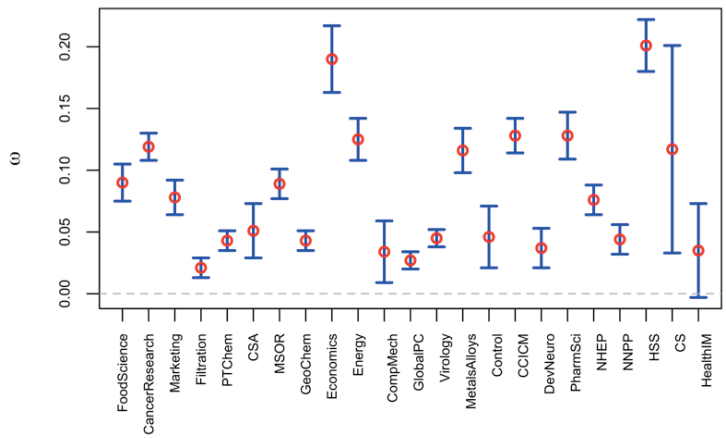


Figure 2. Zero-modification parameter parameters and 95% confidence intervals relative to a zero-modified discretised log-normal distribution for 23 Scopus categories

Zero-modified hooked power law distribution

Relative to a hooked power law distribution both significant positive (13 subjects) and significant negative (6 subjects) estimates of the zero-modification parameter occur, as well as 4 non-significant estimates (Figure 3, see also Table A2). There is both zero-inflation and zero-deflation, and possibly no zero-modification relative to the discretised log-normal distribution.

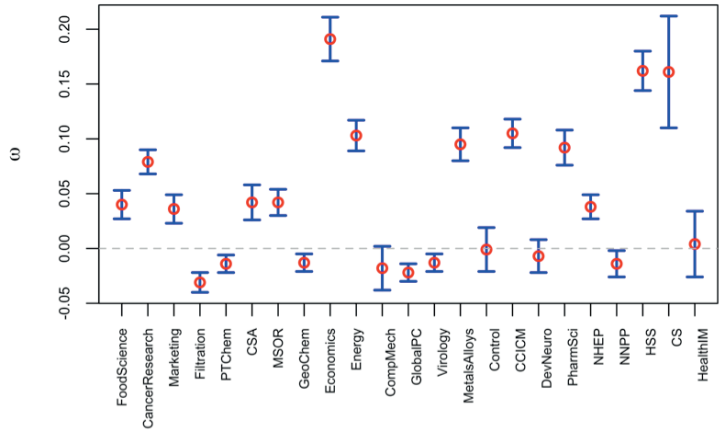


Figure 3. Zero-modification parameters and their confidence intervals relative to a zero-modified hooked power law distribution for 23 Scopus categories

Zero-modified discretised Weibull distribution

Relative to a discretise Weibull distribution only one estimate of the zero-modification parameter is significantly positive, 15 being significantly negative and 7 non-significant (Figure 4, see also Table A3). There is both zero-inflation and zero-deflation and possibly no zero-modification relative to the discretise Weibull distribution.

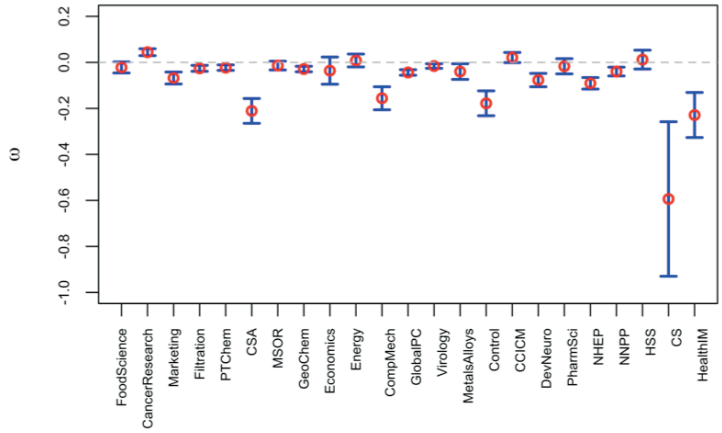


Figure 4. Zero-modification parameters and their confidence intervals relative to a zero-modified discretise Weibull distribution for 23 Scopus categories

Discussion

The results comparing distributions are limited by considering only one year and a small sample of Scopus categories. Other years and results may well give different results. The citation count distributions may also be affected by articles published in January having almost a year longer to be cited than articles published in December. Moreover, the analysis does not take into account factors that influence citation counts, such as individual, institutional and international collaboration; journal and reference impacts; abstract readability; reference and keyword totals; paper, abstract and title lengths.

It is clear from the results that zero-modification occurs relative to a given distribution. For example, the estimated value of the zero-inflation parameter for Neuropsychology and Physiological Psychology is 0.044 relative to the discretised log-normal distribution, but −0.040 relative to a hooked power law distribution, both estimates being significant. Thus, with the former distribution as the base-model there is statistical evidence of zero-inflation and hence "unciteable articles" within the field, but with the latter as the base distribution there is no such evidence of unciteable articles; instead there is evidence that some uncited articles may have been excluded. It is thus important to determine the best fitting base distribution to accurately determine the presence of zero-inflation of zero-deflation (or the absence of either), and the presence of zero inflation/deflation for one model is insufficient to prove that there are perfect or omitted zeros.

The zero-modified hooked power law distribution is the best fitting model for 13 subject areas, the zero-modified Weibull fitting best for 6 subject areas, the other 4 being best fitted by the zero-modified discrete log-normal (Table 1). Thus, whilst (as in Thelwall, 2016), the zero-modified hook power law distribution is still the best fitting model in the majority of cases, the zero-modified discrete Weibull distribution is also a good candidate in some cases. This indicates that either citation counts are best modelled by a single universal distribution that has not yet been considered, or that it is inadvisable to attempt to model citation counts without incorporating predictors, for example, number of authors, into the analyses.

Table 1. AIC values for zero-modified and standard versions of discretised log-normal, hooked power law and Weibull for 23 Scopus categories. Best fitting distributions are in bold.

Subjects	AIC ZMDLN	AIC ZMHPL	AIC ZMWeibull	AIC DLN	AIC HPL	AIC Weibull
Food Science	31448.58	**31393.14**	31398.72	31556.66	31428.84	**31402.40**
Cancer Research	37461.12	**37425.02**	37534.54	37914.00	37687.96	**37564.80**
Marketing	32366.34	**32348.92**	32451.12	32462.34	**32381.32**	32484.36
Filtration & Separation	**13469.50**	13535.42	13553.32	**13507.82**	13566.04	13569.94
Physical & Theoretical Chemistry	36176.02	**36155.10**	36200.76	36304.14	**36166.10**	36216.52
Computer Science Application	**31554.04**	31562.34	31594.68	**31571.44**	31585.56	31707.74
Management Science & Operations Research	34599.58	**34576.70**	34634.12	34771.90	**34632.34**	34636.38
GeoChemistry & Petrology	36667.80	**36661.82**	36738.60	36789.84	**36670.72**	36762.96
Economics & Econometrics	26953.04	26970.90	**26933.90**	27067.86	27195.56	**26935.50**
Energy Engineering & Power Technology	31841.10	31813.28	**31748.70**	31982.44	32000.72	**31749.06**
Computational Mechanics	15564.98	**15559.42**	15591.68	15571.52	**15562.38**	15654.46
Global & Planetary Change	**29959.92**	29986.90	30089.84	30019.14	**30012.50**	30145.58
Virology	**37266.22**	37369.72	37448.60	37471.16	**37380.06**	37457.88
Metals & Alloys	31544.04	31526.70	**31512.86**	31650.52	31671.40	**31519.04**
Control & Optimization	18375.46	**18369.88**	18420.50	18386.84	**18369.90**	18493.82
Critical Care & Intensive Care Medicine	35467.04	35432.32	**35428.70**	35706.52	35726.20	**35431.84**
Developmental Neuroscience	14550.92	**14526.48**	14579.18	14570.48	**14527.44**	14615.60
Pharmaceutical Science	29745.64	29721.02	**29715.20**	29864.16	29844.84	**29716.36**
Nuclear & High Energy Physics	35596.62	**35554.86**	35785.94	35716.84	**35600.62**	35856.58
Neuropsychology & Physiological Psych	20236.10	**20219.00**	20266.78	20289.52	**20224.16**	20287.24
Health Social Science	27138.34	**27117.98**	27183.22	27324.98	27372.70	**27183.54**
Cultural Studies	16745.46	16750.88	**16744.14**	**16751.40**	16774.10	16777.56
Health Information Management	6772.300	**6768.460**	6811.520	6775.280	**6768.540**	6854.220

Conclusion

This article introduced the zero-modified hooked power law, discrete log-normal and Weibull distributions. The new method also allows the estimation of both positive and negative zero-modification parameters, enabling the determination of confidence intervals for and statistical tests of parameter estimates. The results showed that each distribution fits citation count data better than the others for some Scopus categories. The results also show that both zero-inflation and zero-deflation occur for citation count data, but changing a base model can alter one type to another. As a consequence of this, it is important to be wary of making definitive statements concerning zero-inflation or zero-deflation.

References

Akaike, H. (1974). A new look at the statistical model identification. IEEE Transactions on Automatic Control, 19(6):716–723.

Brzezinski, M. (2015). Power laws in citation distributions: evidence from scopus. Scientometrics, 103(1):213–228.

Dietz, E. & Böhning, D. (2000). On estimation of the Poisson parameter in zero-modified Poisson models. Computational Statistics & Data Analysis, 34(4):441 – 459.

Faraway, J. J. (2005). Extending the linear model with r. generalized linear, mixed effects and nonparametric regression models. Chapmen Hall, Boca Raton, FL, USA, ix.

Greenland, S., Senn, S., Rothman, K., Carlin, J., Poole, C., Goodman, S., & Altman, D. (2016). Statistical tests, p values, confidence intervals, and power: a guide to misinterpretations. European Journal of Epidemiology, 31(4):337–350.

Mendonca, L. (1995). Longitudinalstudie zu kariespraventiven methoden, durchgefuhrt bei 7 bis 10 jahrigen urbanen kindern in belo horizonte (brasilien). Inaugural-Dissertation zur Erlangung der zahnmedizinischen Doktorwurde am Fachbereich Zahn, Mund und Kieferheilkunde der Freien Universitat Berlin.

Pennock, D. M., Flake, G. W., Lawrence, S., Glover, E. J., & Giles, C. L. (2002). Winners don't take all: Characterizing the competition for links on the web. PNAS, 99(8):5207–5211.

Thelwall, M. (2016). Are there too many uncited articles? zero-inflated variants of the discretised lognormal and hooked power law distributions. Journal of Informetrics, 10(2):622 – 633.

Vinciotti, V. (2016). DWreg: Parametric Regression for Discrete Response. R package version 2.0.

Wasserman, L. (2006). All of Nonparametric Statistics (Springer Texts in Statistics). Springer-Verlag, Berlin, Heidelberg.

Wilson, P. (2015). The misuse of the Vuong test for non-nested models to test for zero-inflation. Economics Letters, 127(C):51–53.

Wilson, P. & Einbeck, J. (2018). A new and intuitive test for zero-modification. Statistical Modelling. Available Online.https://doi.org/10.1177/1471082X18762277.

Wasserstein, R. L. and Lazar, N. A. (2016). The asa's statement on p-values: Context, process, and purpose. The American Statistician, 70(2):129–133.

Appendix: Extra details of the parameter estimates.

Table A1. Parameter estimates: Zero-modified discretised log-normal for 23 Scopus categories

Subjects	ω	SE_{ω}	CIL_{ω}	CLR_{ω}	p-value	μ	σ
Food Science	0.090	0.0075	0.075	0.105	0.00000	1.82	1.02
Cancer Research	0.119	0.0054	0.108	0.13	0.00000	2.43	1.06
Marketing	0.078	0.0071	0.064	0.092	0.00000	1.98	1.10
Filtration & Separation	0.021	0.0042	0.013	0.029	0.00000	2.58	0.90
Physical & Theoretical Chemistry	0.043	0.0040	0.035	0.051	0.00000	2.28	0.94
Computer Science Application	0.051	0.0114	0.029	0.073	0.00001	1.56	1.29
Management Science & Operations Research	0.089	0.0061	0.077	0.101	0.00000	2.10	1.05
GeoChemistry & Petrology	0.043	0.0040	0.035	0.051	0.00000	2.31	0.96
Economics & Econometrics	0.190	0.0136	0.163	0.217	0.00000	1.34	1.29
Energy Engineering & Power Technology	0.125	0.0087	0.108	0.142	0.00000	1.82	1.15
Computational Mechanics	0.034	0.0126	0.009	0.059	0.00697	1.65	1.06
Global & Planetary Change	0.027	0.0037	0.020	0.034	0.00000	2.49	1.00
Virology	0.045	0.0035	0.038	0.052	0.00000	2.41	0.91
Metals & Alloys	0.116	0.0094	0.098	0.134	0.00000	1.75	1.18
Control & Optimization	0.046	0.0129	0.021	0.071	0.00036	1.57	1.08
Critical Care & Intensive Care Medicine	0.128	0.0070	0.114	0.142	0.00000	2.16	1.20
Developmental Neuroscience	0.037	0.0081	0.021	0.053	0.00000	2.18	1.07
Pharmaceutical Science	0.128	0.0096	0.109	0.147	0.00000	1.66	1.09
Nuclear & High Energy Physics	0.076	0.0063	0.064	0.088	0.00000	2.13	1.13
Neuropsychology & Physiological Psych	0.044	0.0059	0.032	0.056	0.00000	2.21	0.98
Health Social Science	0.201	0.0108	0.180	0.222	0.00000	1.54	1.07
Cultural Studies	0.117	0.0431	0.033	0.201	0.00664	0.10	1.20
Health Information Management	0.035	0.0193	-0.003	0.073	0.06976	1.83	1.24

Table A2. Parameter estimates: Zero-modified hooked power law for 23 Scopus categories

Subjects	ω	SE_{ω}	CIL_{ω}	CLR_{ω}	p-value	μ	σ
Food Science	0.040	0.0067	0.027	0.053	0.00000	40.21	6.37
Cancer Research	0.079	0.0055	0.068	0.090	0.00000	61.35	5.34
Marketing	0.036	0.0064	0.023	0.049	0.00000	27.28	4.18
Filtration & Separation	-0.031	0.0048	-0.040	-0.022	0.00000	190.92	12.52
Physical & Theoretical Chemistry	-0.014	0.0042	-0.022	-0.006	0.00086	121.18	10.90
Computer Science Application	0.042	0.0084	0.026	0.058	0.00000	11.31	2.96
Management Science & Operations Research	0.042	0.0059	0.030	0.054	0.00000	40.78	5.15
GeoChemistry & Petrology	-0.013	0.0042	-0.021	-0.005	0.00197	93.37	8.47
Economics & Econometrics	0.191	0.0101	0.171	0.211	0.00000	10.05	3.06
Energy Engineering & Power Technology	0.103	0.0073	0.089	0.117	0.00000	26.56	4.38
Computational Mechanics	-0.018	0.0103	-0.038	0.002	0.08054	20.28	4.39
Global & Planetary Change	-0.022	0.0039	-0.030	-0.014	0.00000	79.88	6.37
Virology	-0.013	0.0039	-0.021	-0.005	0.00086	127.17	10.26
Metals & Alloys	0.095	0.0076	0.080	0.110	0.00000	19.76	3.77
Control & Optimization	-0.001	0.0102	-0.021	0.019	0.92190	17.34	4.14
Critical Care & Intensive Care Medicine	0.105	0.0064	0.092	0.118	0.00000	31.22	3.86
Developmental Neuroscience	-0.007	0.0075	-0.022	0.008	0.35065	45.73	5.21
Pharmaceutical Science	0.092	0.0080	0.076	0.108	0.00000	23.23	4.66
Nuclear & High Energy Physics	0.038	0.0058	0.027	0.049	0.00000	31.67	4.15
Neuropsychology & Physiological Psych	-0.014	0.0060	-0.026	-0.002	0.01963	70.08	7.30
Health Social Science	0.162	0.0091	0.144	0.180	0.00000	17.89	4.33
Cultural Studies	0.161	0.0260	0.110	0.212	0.00000	4.25	3.47
Health Information Management	0.004	0.0155	-0.026	0.034	0.79636	13.47	2.94

Table A3. Parameter estimates: Zero-modified Weibull for 23 Scopus categories

Subjects	ω	SE_{ω}	CIL_{ω}	CLR_{ω}	p-value	μ	σ
Food Science	-0.022	0.0121	-0.046	0.002	0.06904	0.82	0.80
Cancer Research	0.044	0.0076	0.029	0.059	0.00000	0.90	0.81
Marketing	-0.068	0.0133	-0.094	-0.042	0.00000	0.80	0.70
Filtration & Separation	-0.026	0.0064	-0.039	-0.013	0.00005	0.95	1.00
Physical & Theoretical Chemistry	-0.023	0.0059	-0.035	-0.011	0.00010	0.91	0.94
Computer Science Application	-0.211	0.0273	-0.265	-0.157	0.00000	0.67	0.56
Management Science & Operations Research	-0.014	0.0098	-0.033	0.005	0.15313	0.85	0.78
GeoChemistry & Petrology	-0.029	0.0062	-0.041	-0.017	0.00000	0.91	0.90
Economics & Econometrics	-0.036	0.0303	-0.095	0.023	0.23479	0.64	0.57
Energy Engineering & Power Technology	0.008	0.0142	-0.020	0.036	0.57318	0.80	0.72
Computational Mechanics	-0.156	0.0256	-0.206	-0.106	0.00000	0.74	0.70
Global & Planetary Change	-0.044	0.0063	-0.056	-0.032	0.00000	0.91	0.86
Virology	-0.016	0.0052	-0.026	-0.006	0.00209	0.93	0.95
Metals & Alloys	-0.040	0.0172	-0.074	-0.006	0.02004	0.76	0.66
Control & Optimization	-0.178	0.0275	-0.232	-0.124	0.00000	0.71	0.66
Critical Care & Intensive Care Medicine	0.021	0.0113	-0.001	0.043	0.06311	0.83	0.70
Developmental Neuroscience	-0.077	0.0146	-0.106	-0.048	0.00000	0.85	0.76
Pharmaceutical Science	-0.017	0.0167	-0.050	0.016	0.30870	0.76	0.72
Nuclear & High Energy Physics	-0.091	0.0126	-0.116	-0.066	0.00000	0.80	0.66
Neuropsychology & Physiological Psych	-0.040	0.0095	-0.059	-0.021	0.00003	0.89	0.87
Health Social Science	0.012	0.0211	-0.029	0.053	0.56955	0.70	0.67
Cultural Studies	-0.594	0.1712	-0.930	-0.258	0.00052	0.30	0.49
Health Information Management	-0.229	0.0501	-0.327	-0.131	0.00000	0.70	0.56

Teams Prevent Misconduct: A Study of Retracted Articles from the Web of Science

Justus M. K. Rathmann[1] and Heiko Rauhut[2]

[1] *rathmann@soziologie.uzh.ch*
Institute of Sociology, University of Zurich, Andreasstr. 15, CH-8050 Zurich, Switzerland

[2] *rauhut@soziolgie.uzh.ch*
Institute of Sociology, University of Zurich, Andreasstr. 15, CH-8050 Zurich, Switzerland

Abstract

Collaborations become increasingly important among almost all scientific disciplines. Teams can be more productive and achieve more attention to their work. It is, however, unclear, whether teams also lead to higher research integrity. On the one hand, there may be a volunteer's dilemma in larger groups such that the responsibility diffuses who is controlling whom. The 'volunteer hypothesis' predicts that the more co-authors, the more scientific misconduct. On the other hand, larger groups may also achieve a higher level of social control. The 'control hypothesis' predicts that the more co-authors, the less scientific misconduct. Retractions are used as an operationalization of scientific misconduct. The data collection comprises of retracted articles from the Web of Science data set. In addition, control groups of non-retracted articles are constructed, using methods known from causal inference and bibliometrics. The analyses demonstrate that larger author groups have a lower retraction probability compared to smaller author groups. This suggests that teams prevent misconduct, most likely by their higher ability to social control. The results indicate that the development towards more and larger research collaborations may have positive macro-level consequences for the system of science.

Introduction

Collaboration is an emerging topic in science studies (Hall et al., 2018). The number and share of research in teams is steadily increasing. At the same time, teams become larger. These findings are robust across the scientific disciplines. There are good reasons for such developments; collaborating scientists can share their work and concentrate on their strengths, additionally, collaborating with others can generate new ideas. In addition, scientific articles with a larger number of co-authors are cited more often than articles with fewer co-authors (Wuchty, Jones & Uzzi, 2007). However, it remains an open question, whether teams also lead to higher research integrity.

The Office of Research Integrity (ORI) identifies research misconduct as consisting of data fabrication, data falsification and plagiarism and underlines that honest errors or differences of opinions are not misconduct (The Office of Research Integrity, 2011). One way of uncovering scientific misconduct and inherently flawed research are so-called retractions (Hesselmann et al., 2016). This is the case because only a small fraction of retractions is due to honest mistakes; the vast majority of retractions are due to various forms of scientific misconduct (Fang, Steen & Casadevall, 2012; Gewinner, Diekmann & Rathmann, u.r.).

Hesselmann et al. (2016) summarize the state of the art on retractions; compared to other publication types, retractions are very uncommon, representing only 0.02% of the PubMed database. However, the share of retractions has been increasing steadily since the end of the last century, this growth has accelerated since the early 2000s. In absolute terms, most retractions come from the United States, whereas the emerging science nations generate the largest share of retractions on total scientific output. Retractions are most prevalent in medicine, chemistry, life sciences and multidisciplinary studies. Articles in high-impact journals and highly cited articles are retracted comparatively more frequently.

Studies on the relationship between the number of co-authors and retractions have yielded mixed results so far. In addition, these investigations have mostly referred to studies in bio-medicine, medicine and life sciences. Foo (2010) found that only 6.6\% of all retractions in bio-

medicine and life sciences fall on individual authors. In these disciplines, however, single authorship is uncommon. Furman, Jensen & Murray (2012) have found no statistical evidence of the number of co-authors affecting the likelihood of retraction. However, other research concluded that fraudulent retractions had significantly more co-authors than erroneous retractions (Steen, 2010) and that randomized clinical trials with fewer authors had a higher likelihood of retraction (Steen & Hamer, 2014). This leads to the following research question: Does the size of the research team influence the probability of an article being retracted?

This research question will be addressed using publication data from the Web of Science (WoS). Using a matching procedure, the so-called bibliographic coupling (Kessler, 1963), retracted and ordinary research articles will be sampled into a control and a treatment group, resulting in a quasi-experimental design. The treatment is the retraction process. As a consequence, differences in the number of authors in the two groups can be attributed to the behaviour underlying the treatment.

Theoretical Considerations

Two contradictory hypotheses can be derived from theories of actor models and group processes; 'social control' on the one hand and the 'diffusion of responsibility' on the other hand. In group contexts, people inform themselves about context-specific group norms by observing identity-consistent behaviour of core group members in order to adapt and integrate (Hogg et al., 2004). Good scientific practice is an important norm in science (The Office of Research Integrity, 2011). The more co-authors, the higher the probability that one of them has a sceptical attitude and demands receiving in-depth information on the production process of critical matters. Auspurg & Hinz (2011) show, for instance, that articles by single authors are more prone to publication bias than articles by multiple co-authors. This leads to the first hypothesis:

H_a: The larger the number of co-authors of an article, the lower is the likelihood of retraction.

The concept of 'diffusion of responsibility' predicts that the probability of volunteering to do something negatively correlates with group size (Darley & Latane, 1968). Based on this, Diekmann (1985) has developed the so-called 'volunteer's dilemma'. In this game one volunteer is sufficient to provide the public good, which is, however, costly to produce. Both the free-riders and the volunteer profit equally from the public good. The game has the interesting macro-level implication that volunteering becomes less likely the larger the group is. Translated into the scientific context, this means that only one volunteer is needed to review the data and analysis. Hence, the more authors, the higher the probability that all involved think that the others closely checked data collection or processed results, while in fact, nobody may have done it. This leads to the second hypothesis:

H_b: The larger the number of co-authors of an article, the higher is the likelihood of retraction.

Methodological Considerations

In a first step, all retracted articles from the WoS are sampled. There is no ‘retraction’-category in WoS, therefore the retractions were identified by querying for articles with the document type ‘correction’ or ‘correction, addition’ and the word ‘retract’ in the title. These serve in the later analysis as a treatment group. Subsequently, four control groups are formed on different levels of similarity.

The control groups and their matching mechanism are presented in Table 1. The first control group (CG Field) is based on a random discipline sampling. The retractions are matched with a random article from the same WoS subject category. It is important to note that an article in the WoS can have several subject categories, which are assigned based on the subject categories of the journal. The second control group (CG Journal) is based on random journal sampling. Articles from the same journal and year are matched with the retracted articles.

The third and fourth control groups are determined on the basis of the bibliographic couplings. In this matching mechanism, the similarity of articles is determined on the basis of shared references; the more shared references between two articles, the more similar the articles are (Kessler, 1963). Articles by an author with the same name as the author of a retracted article are rejected as potential matches in order to avoid to match the author with himself. In case of a tie, one article is randomly selected from the best matches. The third control group (CG Coupling) is, thus, based on regular bibliographic coupling. Since the number of references of an article positively correlates with the number of co-authors (Glänzel & Schubert, 2004), an improved bibliographic coupling algorithm is introduced. The number of shared references, is in this algorithm divided by the number of co-authors to control for this correlation. This matching procedure is used for the fourth control group (CG Imp. Coupling).

Table 1. Control groups and measures of similarity

Control Group	*Matching Partner*
CG Field	Random article from the same field and year
CG Journal	Random article from the same journal and year
CG Coupling	Article with the highest number of shared references
CG Imp. Coupling	Article with the highest number of shared references divided by the number of co-authors

Figure 1 shows a schematic plot of the bibliographic coupling mechanism. The retracted article by Miller would be matched with the article by Johnson. The articles have two references in common which is more than the shared number of references by the articles of Perry and Miller. The other article by Miller is excluded.
This methodological structure results in a quasi-experimental design with pairwise-matching and no prior measurement.

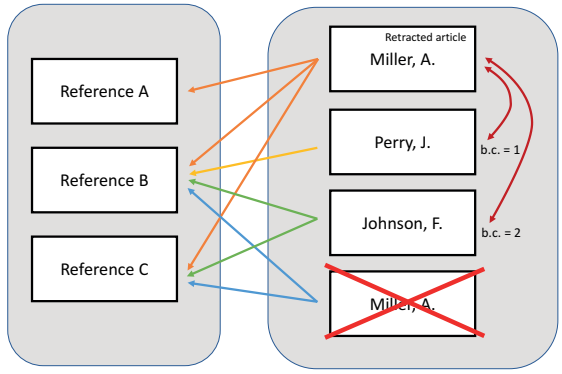


Figure 1. Schematic plot of bibliographic coupling

Preliminary Results

The data set consists of 3,549 retractions which make up the treatment group. CG Field and CG Journal each consist of 3,549 matched articles. The matching algorithms for the bibliographic

coupling are not finished to date, therefore there is no data for CG Coupling and CG Imp. Coupling yet.

For the following analyses, the number of authors of an article is field-normalized. The field-normalization is similar to the *z*-transformation. The field-normalized number of authors A_{norm} is obtained by subtracting the average number of authors in the field μ_{field} from the number of authors A and dividing the result by the standard deviation of the number of authors in the field σ_{field} (see equation below). The field is determined by the subject category in the WoS database.

$$A_{\text{norm}} = \frac{A - \mu_{\text{field}}}{\sigma_{\text{field}}}$$

T-tests are used to compare the field-normalized average number of authors in the treatment group with the number of authors in CG field and CG journal. The mean value of the treatment group is -0.33. This means that retracted articles have 0.33 standard deviations fewer authors than it is usual in the respective field. The mean value of CG field is -0.02, the t-test yields a *t*-value of 14.8. The mean value of CG journal is 0.09, the t-test yields a *t*-value of 15.2. The difference for both control groups is therefore significant at $p < 0.01$. Retracted articles, thus, have significantly fewer authors than randomly matched articles from the same field and the same journal.

Additionally, logistic regression models with the variable *retraction* as the dependent variable are conducted. This variable is dichotomous, 1 for retracted articles and 0 for regular articles. The detailed regression tables are presented in Table 2 in the appendix. The number of authors is modelled by a second order polynomial function to model a non-linear relationship. The predicted retraction probabilities are depicted in Figure 2 and figure 3. Since each sample consists of 50% retracted and 50% regular articles the baseline retraction probability is 50%.

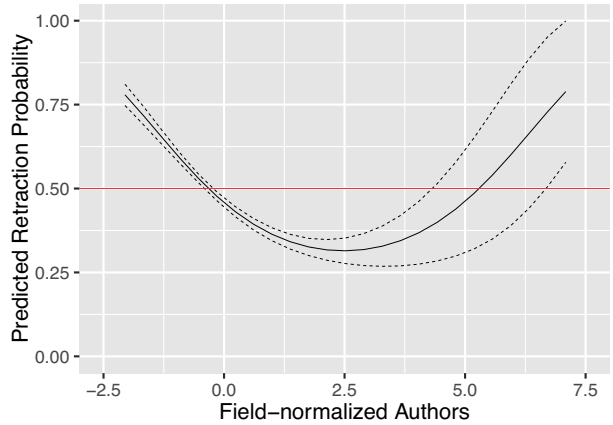


Figure 2. Predicted retraction probability given field-normalized authors (CG Field)

Figure 2 plots the predicted regression retraction probabilities in the random field sample. The predicted retraction probability decreases steadily as the number of authors increases until reaching 2.5 standard deviations. Therefore, the function increases steadily resulting in a u-shaped function. Thus, articles from author groups smaller than usual in the field are more likely to be retracted. The same effect can be found for very large author groups with more than 5 standard deviation from the field mean.

Figure 3 plots the predicted regression retraction probabilities in the random journal sample. As in the field sample, the predicted retraction probability continuously decreases as the number of authors increases up to 2.5 standard deviations. Unlike in the field sample, the predicted

retraction probability in the journal sample decreases before reaching a plateau at 5 standard deviations. This sample also shows that articles of groups whose size is below average have a higher probability to become retracted. This pattern holds for a wide range of author group sizes.

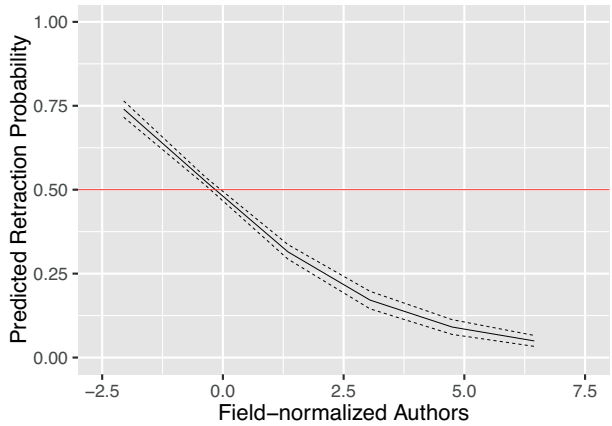


Figure 3. Predicted retraction probability given field-normalized authors (CG Journal)

Preliminary Conclusion

Retracted articles are more likely to originate from author groups which are smaller than usual in the respective field. This result can be shown in both t-tests and logistic regressions in different samples. These preliminary results therefore discard the idea of a diffusion of responsibility developed in hypothesis H_b and indicate a case for social control as stated in hypothesis H_a. Albeit, the diffusion of responsibility might play a role for very large author groups. However, one can also argue for a selection effect; if a researcher knows he wants to perpetrate scientific misconduct, he might choose to work alone or in a small team in order to have as few accessories as possible. These findings are backed up by (preliminary) findings in other areas of misconduct; Auspurg & Hinz (2011) find publication bias more likely to be present in articles by single authors, Horbach & Halffman (2019) present evidence that articles containing problematic text recycling have less authors than articles not containing problematic recycling. Although the results so far are robust, it is not yet possible to draw a final conclusion. The bibliographic coupling allows a significant step forward in the determination of similarity and could lead to new results.

In all matching processes, there is a chance to match a retracted article with an article that will be retracted in future. It can take as long as 35 months for an article to be retracted (Decullier, Huot & Maisonneuve, 2014). However, this chance is very small due to the low prevalence of retractions. Furthermore, relatively few authors are responsible for a large part of the retractions (Gewinner, Diekmann & Rathmann, u.r.) and this bibliographic coupling algorithm excludes matching an author with itself.

Acknowledgements

The Authors thank Antonia Velicu for research assistance. Justus Rathmann is funded by the Swiss National Science Foundation (SNSF) under the SNSF Starting Grant "CONCISE" (S-64408-01-01).

References

Auspurg, K & Hinz, T. (2001) What fuels publication bias? *Jahrbücher für Nationalökonomie und Statistik*, 231(5-6).

Darley, J.M. & Latane, B (1968). Bystander intervention in emergencies: Diffusion of responsibility. *Journal of Personality and Social Psychology* 8(4), 377-383.

Diekmann, A. (1985). Volunteer's dilemma. *Journal of Conflict Resolution*, 29(4), 605-610.

Decullier, E., Huot, L. & Maisonneuve, H. (2014). What time-lag for a retraction search of PubMed? *BMC Resarch Notes*, 25(7), 395.

Fang, F.C., Steen, R.G. & Casadevall, A. (2012). Misconduct accounts for the majority of retracted scientific publications. *Proceedings of the National Academy of Sciences*, 109(42), 17028-17033.

Foo, J.Y.A. (2010). A retrospective analysis of the trend of retracted publications in the field of biomedical and life sciences. *Science and Engineering Ethics*, 17(3), 459-468.

Furman, J.L., Jensen, K. & Murray, F. (2012). Governing knowledge in the scientific community: Exploring the role of retractions in biomedicine. *Research Policy*, 41(2), 276-290.

Gewinner, J., Diekmann, A., & Rathmann, J.M.K. (under review). Why are Articles Retracted?
A Systematic Analysis of the Reasons Behind Retraction from the 'Retraction Watch' Blog. *Science, Technology, & Human Values*.

Glänzel, W. & Schubert, A. (2004). Analysing scientific networks through co-authorship. In H.F. Moed, W. Glänzel & U. Schmoch (Eds.), Ha*ndbook of Quantitative Science and Technology Research* (pp. 257-276). Kluwer Academic Publishers.

Hall, K.L., Vogel, A.L., Huang, G.C., Serrano, K.J., Rice, E.L., Tsakraklides, S.P. & Fiore, S.M. (2018). The science of team science: A review of the empirical evidence and research gaps on collaboration in science. *American Psychologist*, 73(4), 532-548.

Hesselmann, F., Graf, V., Schmidt, M. & Reinhart, M. (2016). The visibility of scientific misconduct: A review of the literature on retracted journal articles. *Current Sociology*, 65(6). 814-845.

Hogg, M.A., Abrams, D., Otten, S. & Hinkle, S. (2004). The social identity perspective. *Small Group Research*, 35(3), 246-276.

Horbach, S.P.J.M. & Halffman, W. (2019). The extent and cause of academic text recycling or self-plagiarism. *Research Policy*, 48(2), 492-502.

Kessler, M.M. (1963). Bibliographic coupling between scientific papers. *American Documentation*, 14(1), 10-25.

Steen, R.G. & Hamer, R. (2014). A case-control comparison of retracted and non-retracted clinical trials: Can retraction be predicted? *Publications*, 2(1), 27-37.

Steen, R.G. (2010). Retractions in the scientific literature: Is the incidence of research fraud increasing? *Journal of Medical Ethics*, 37(4), 249-253.

The Office of Research Integrity. (2011). *Definition of research misconduct*. Retrieved 25 January 2019 from: https://ori.hhs.gov/definition-misconduct.

Wuchty, S., Jones, F.B. & Uzzi, B. (2007). The increasing dominance of teams in the production of knowledge. *Science*, 316(5827), 1036-1039.

Appendix

Table 2. Retractions compared to the average number of authors in the field and journal

	Field sample	Field sample	Journal sample	Journal sample
Authors	-0.414***	-0.492***	-0.511***	-0.527***
	(0.029)	*(0.031)*	*(0.030)*	*(0.030)*
Authors^2		0.099***		0.013***
		(0.014)		*(0.001)*
Intercept	-0.077***	-0.168***	-0.078***	-0.091***
	(0.024)	*(0.028)*	*(0.025)*	*(0.025)*
N	7,098	7,098	7,098	7,098
AIC	9,623	9,578	9,496	9,486

* $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$; S.E. in parenthesis

A two-step approach toward subject prediction

Shenghui Wang[1] and Rob Koopman[2]
[1] shenghui.wang@oclc.org
[2] rob.koopman@oclc.org
OCLC Research, Schipholweg 99, 2316XA Leiden, The Netherlands

Abstract

Automatic subject prediction is a desirable feature for modern digital library systems, as manual indexing could no longer cope with the rapid growth of digital collections. Data sparsity and model scalability are the major challenges to solving this extreme multi-label classification problem automatically. In this research-in-progress paper, we propose to address this problem using a two-step approach. We first propose to use an efficient and effective embedding method that embed terms, subjects and documents into the same semantic space, where similarity could be computed easily. We then describe a novel Non-Parametric Subject Prediction (NPSP) method and show how effectively it predicts even very specialised subjects, which are associated with few documents in the training set and are more problematic for a classifier.

Introduction

Because of the ever-increasing number of documents that information systems deal with, automatic subject indexing, i.e., identifying and describing the subject(s) of documents to increase their findability, is one of the most desirable features for many such systems. Subject index terms are normally taken from knowledge organization systems (e.g., thesauri, subject headings systems) and classification systems (e.g., Dewey decimal classification) which easily contain tens or hundreds of thousands of terms or codes. Automatically assigning a small set of relevant subjects from the huge label space -- the Extreme Multi-label Text Classification (XMTC) problem -- is very difficult. Data sparsity and scalability are the major challenges.

In this research-in-progress paper, we describe our two-step approach to addressing this problem. First, we propose to use a novel embedding method which extends random projection by weighting and projecting raw term embeddings orthogonally to an average language vector, thus improving the discriminating power of resulting term embeddings, and build more meaningful document embeddings by assigning appropriate weights to individual terms. Subjects are treated as special terms which get embedded into the same semantic space where terms and documents live. Secondly, we propose a novel Non-Parametric Subject Prediction (NPSP) method to predict subjects for unseen documents. We compare this method with the state-of-the-art deep learning method and the direct subject-document-similarity based method.

Related Work

The goal of automatic subject prediction is to assign a document a small subset of relevant subject labels from tens or hundreds of thousands target labels. This remains a difficult problem and is a form of Extreme Multi-label Text Classification (XMTC) (Prabhu and Varma, 2014; Bhatia et al., 2015; Liu et al., 2017), where the prediction space normally consists of hundreds of thousands to millions of labels. Data sparsity and scalability are the major challenges. Different from traditional binary or multi-class classification problems, this XMTC problem cannot assume that the target labels are independent or mutually exclusive. Scalable solutions became available only in recent years (Bhatia et al., 2015; Prabhu and Varma, 2014). There are four categories of solutions: 1) 1-vs-All (Prabhu et al., 2018), 2) Embedding- (Bhatia et al., 2015), 3) Tree-based (Prabhu and Varma, 2014), and 4) Deep learning methods (Joulin et al., 2016; Liu et al., 2017).

Method

We propose to embed subjects, terms and documents in a single semantic space. This allows us to find a vector representation of a document and to compute its similarity to the vector representation of a subject, thus allowing us to use similarity between a query document and

the subject candidates to find the most appropriate subjects for that document. In addition, we propose computing similarities between the query document and previously seen documents to better assess the validity of a subject to the query document.

Ariadne semantic embedding

Let a document be a set of terms (words or 2-word phrases) for which term co-occurrence is relevant and which can be meaningfully annotated with a subject. In general, a document could therefore be a sentence, a paragraph, a fixed-size window, or, in our case, a bibliographic record. Let n_D be the total number of documents, n_S the number of *frequent* subjects, n_V the number of *frequent* terms. A term or subject is considered frequent when it occurs in more than K documents in the corpus, where K is flexible depending on the size of the corpus. In addition, let n_E be the total number of entities – which could be terms, subjects, authors, citations, journals – we want to embed, and D the chosen dimensionality of the embedding vectors.

Building on our previous work (Koopman et. al, 2015, 2017 and 2019) we embed the relevant entities by Random Projection (Achlioptas, 2003; Johnson and Lindenstrauss, 1984) of their weighted co-occurrences:

$$C'_{n_E \times D} = C_{[n_E \times n_S]} R_{[n_S \times D]}$$

where C' is the matrix of embedding vectors, C is the co-occurrence matrix of different terms and R is a random matrix. In this work, we focus on subjects, and in this particular use case we observe that it is useful to use co-occurrence of the entities with the subject labels only. In general, term-term co-occurrences are more common and well-suited, in which case we would haven n_V columns to the matrix C and n_V rows to R.

Orthogonal projection

Traditional models discard both very infrequent words (because they are too rare for the model to be able to capture their semantics from the training data) and very frequent words (so-called "stop words" because they do not provide any semantically useful information). In our approach, we compute the average "language vector" of the corpus, $\overrightarrow{v_a}$, the sum of all the rows of C'. Unsurprisingly, this vector is very similar to the average vector of stop words. Intuitively, words are increasingly more informative as they differ more from the average vector. By this reasoning, we project word vectors on the orthogonal hyperplane to $\overrightarrow{v_a}$:

$$\overrightarrow{v_t^*} = \overrightarrow{v_t} - (\overrightarrow{v_t} \cdot \overrightarrow{v_a})\, \overrightarrow{v_a} \ ,$$

resulting in a representation where the uninformative component of terms is eliminated, and normalise the vectors to have unit length.

Weight assignment

Using the projection described above, the component that differentiates a term from the average vector is kept as its final embedding. Meanwhile, how different a term is from $\overrightarrow{v_a}$ also indicates how much that term contributes to the semantics of a document it is part of. In order to give a higher weight to the most informative terms, we assign a higher weight to words with lower mutual information to $\overrightarrow{v_a}$ by setting the final weight of each term to be: $w_t = 1 - \cos(\overrightarrow{v_t}, \overrightarrow{v_a})$. This step is crucial to get distinctive document embeddings.

Document embedding

With the embeddings of the frequent entities and their proper weights, we can compute document embedding as the weighted average of its component terms' embeddings.

Prediction by subject-document similarity

Once subjects and documents are embedded in the same semantic space, it is straightforward to calculate the similarity between any subject and any document. Our assumption is that an article would be indexed by its most related subject headings, *i.e.* the subject headings with the highest cosine similarities to the article itself.

NPSP: Non-Parametric Subject Prediction

We now propose a non-parametric algorithm for subject prediction. The algorithm returns a ranked list of subjects, where the subjects are sorted according to a summation of 1) the similarity of each subject to the document and 2) the similarity of those of the k most similar documents from the training set which are annotated with the subject. This combination provides us with a robust ranking measure, which combines the direct embedding of the subject in the semantic space where the documents also live and an extra component which lets the k nearest neighbor documents of the new document vouch for the validity of the subject. The idea is that the embedding of each document is more precise than the embedding of the subjects (since that is done based on a combination of many documents), making the similarity computation more trustworthy and the subjects those documents are annotated with reflect more likely to fit the target document.

Dataset and experiments

The ASTRO dataset (available via http://www.topic-challenge.info/) contains bibliographic information of 111,616 articles published between 2003-2010 in 59 Astronomy and Astrophysics journals indexed by the Web of Science and assigned by the Journal Citation Report to the Astronomy and Astrophysics subject field. This data set was split into the training set (containing 102,869 articles) and the testing set (containing 5455 articles). In the training set, each article has a title, an abstract, a journal ISSN, in average 7.6 authors, 39.5 citations and 10.1 subjects. Different types of entities including terms extracted from the titles and abstracts, ISSNs, authors, citations and subjects were all embedded using Random Projection based on their co-occurrences with the frequent subjects. Infrequent entities that occur in less than 10 articles were discarded. Table 1 lists some stats about these entities.

Table 1. ASTRO dataset stats

Entity	#Total	#Frequent
Subject	93,566	10,624
Author	87,637	17,765
Citation	891,827	88,510
ISSN	59	58
Terms	105,062	143,946

After orthogonal projection, each entity was embedded as a 256-dimensional vector and its weight was assigned based on its similarity to the average language vector as described above. When calculating the embedding for each article, we computed the weighted average of all the entities except subjects as the final embedding. Now different types of entities and articles were embedded in the same semantic space, where their similarity could be computed easily. Each unseen article in the testing set was also embedded into the same semantic space based on its title, abstract, authors and citations. Therefore, its subjects could be predicted using subject-document similarity or the NPSP algorithm. We also applied FastText (Joulin et. al. 2016) which is a state-of-the-art multi-label text classifier to the training set and compared all the predictions using the testing set.

Evaluation

The goal of subject prediction is to provide a shortlist of potentially relevant subjects to describe the document at hand. It is important to present a ranked shortlist of candidate subjects and to evaluate the quality of the prediction with an emphasis on the relevance of the top portion of

such lists. Therefore, we use rank-based evaluation metrics such as precision and recall at top n. Precision@n is the proportion of the predicted subjects in the top n list that are actual subjects of the test document, while Recall@n is the proportion of the correctly predicted subjects over all actual subjects of the test document.

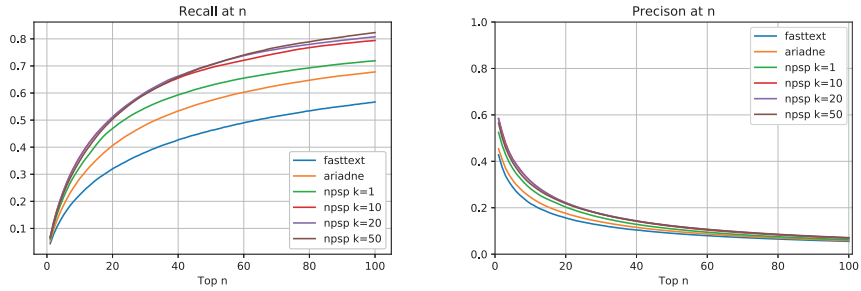


Figure 1. Precision@n and Recall@n

Figure 1 shows the Precision@n and Recall@n of three methods with different settings, where *Ariadne* represents the straightforward predictions based on subject-document similarities. We can see that the precision of the subjects predicted by Ariadne is slightly higher than those generated by FastText, while Ariadne certainly outperforms FastText in terms of recall.

The clear winner is the NPSP method. The precision and recall are both constantly higher than those of the other two methods. Even with $k = 1$, that is, only taking the subjects from the top 1 most similar article, the performance is already improved, especially in terms of recall. With a small $k = 10$, the Precision@1 is 15% higher than FastText, and 13% higher than Ariadne. The improvement in terms of recall is even bigger; the Recall@100 is 22% higher than FastText and 11% higher than Ariadne. The effect of increasing k is not obvious. With $k = 50$, the precision@n does not seem to change, but the recall at the lower ranks has a minimum increase.

A closer look

Table 2 lists the actual subjects of the article, titled "The International DORIS Service (IDS): Toward maturity," and the predictions by the three methods. The first column gives the raw document counts of the actual subjects in the training set. Half of the actual subjects occurred in less than 30 articles in the training set, some of which only occurred in a very few articles or never occurred before. These extremely infrequent subjects are difficult to predict in general. FastText tends to predict common subjects, such as "model" correctly, but "earth" and "system" incorrectly (see the document counts in the last column). Even if predicted correctly, these common subjects are less informative about the article itself. Ariadne successfully predicts more specific infrequent subjects, such as "doris" and "terrestrial reference frame," but misses common ones such as "model." Our NPSP method manages to predict more specific but infrequently subjects as well as the common ones too. This makes the Recall@15 as 50%, and Precision@15 as 46.7%.

We realise that this evaluation has its limitations. As shown in Table 2, highly related subjects such as "geodesy" and "envisat" (Environmental Satellite) are predicted as good candidates for this article, both of which are reasonable and potentially useful, but since they are not the subjects that the human indexer have chosen, their value cannot be easily assessed. This illustrates how precision/recall may not be a very meaningful evaluation metric in this application.

That being said, we believe our predictions are still useful in practice when the predicted subjects are presented to cataloguer as candidate subjects to choose from. A high recall is more important as it would greatly reduce the search space and also provide opportunities for

cataloguers to find more suitable subjects which they probably have not thought of themselves. In the future, we will get subject specialists involved to conduct such qualitative evaluations.

Conclusion

In this research-in-progress paper, we proposed a two-step approach to addressing the problem of subject prediction. We have shown that a similarity-based subject prediction based on a suitable semantic space that allows for the embedding of both terms and subjects is very competitive with the state-of-the-art subject-prediction method based on a classifier. We have described such an embedding and have shown how effective this specific semantic space really is. In addition, we proposed a novel, non-parametric, similarity-based method with the documents instead of the subjects. We have shown that this method substantially improves the quality of the predictions, both in comparison to the state-of-the-art and to the bare similarity-based method. We also showed how our non-parametric method is particularly effective at correctly predicting very specialised subjects, which are associated with few documents in the training set and are more problematic for a classifier.

Our approach is general and not restricted to subject prediction. In the future will explore further to predict other types of entities, such as citations, authors. We will evaluate our method using the multi-label datasets available from the Extreme Classification Repository (Bahtia et. al, 2019) and conduct more human-involved qualitative evaluation.

References

Achlioptas, D. (2003). Database-friendly random projections: Johnson-Lindenstrauss with binary coins. *Journal of Computer and System Sciences* 66(4), 671–687.

Bhatia, K., Jain, H., Kar, P., Varma, M. & Jain P. (2015). Sparse local embeddings for extreme multilabel classification. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett (Eds.), *Advances in Neural Information Processing Systems* 28, pp. 730–738. Curran Associates, Inc.

Foster, D. V. & Grassberger, P. (2011). Lower bounds on mutual information. *Physical review. E, Statistical, nonlinear, and soft matter physics*. 83. 010101.

Johnson, W. & Lindenstrauss, J. (1984). Extensions of Lipschitz mappings into a Hilbert space. *Contemporary Math. 26*, 189–206.

Joulin, A., Grave, E. Bojanowski, P. & Mikolov, T. (2016). Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*: Volume 2, (Short Papers), 427-431.

Koopman, R., Wang, S., Scharnhorst, A. & Englebienne, G. (2015). Ariadne's thread: Interactive navigation in a world of networked information. In *Proceedings of the ACM Conference Extended Abstracts on Human Factors in Computing Systems*. 1833–1838.

Koopman, R., Wang, S., & Scharnhorst, A. (2017). Contextualization of topics: Browsing through the universe of bibliographic information. *Scientometrics* 111, 2 (May 2017), 1119–1139.

Koopman, R., Wang, S. & Englebienne, G. (2019). Fast and Discriminative Semantic Embedding. In *Proceedings of the 13th International Conference on Computational Semantics*. Gothenburg, Sweden. 235–246.

Liu, J., Chang, W.-C., Wu, Y. & Yang, Y. (2017). Deep learning for extreme multi-label text classification. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '17, New York, NY, USA, 115–124.

Prabhu, Y., Kag, A., Harsola, S., Agrawal, R. & Varma, M. (2018, April). Parabel: Partitioned label trees for extreme classification with application to dynamic search advertising. In *Proceedings of the International World Wide Web Conference*.

Prabhu, Y. and M. Varma (2014). FastXML: A fast, accurate and stable tree-classifier for extreme multilabel learning. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York, NY, USA, 263–272.

Kush Bhatia, Kunal Dahiya, Himanshu Jain, Yashoteja Prabhu & Manik Varma. (2019) The Extreme Classification Repository: Multi-label Datasets & Code. http://manikvarma.org/downloads/XC/XMLRepository.html. Accessed on 30 Jan 2019.

Table 2. Comparison between 14 actual subjects versus the top 15 predicted ones by Ariadne, NPSP ($k = 20$), and FastText, where the ones in bold match the actual subjects. The raw document counts of the actual subjects in the training set and those predicted by NPSP and FastText are also given.

d_t	Actual subjects	Ariadne	NPSP (k=20)	d_t	FastText	d_t
27	doris	**doris**	**doris**	25	gps	94
1	geodetic applications	**terrestrial reference frame**	**terrestrial reference frame**	12	earth	587
0	global geodetic observing system	**topex/poseidon**	**topex/poseidon**	22	system	1717
35	information	service	service	23	reference systems	121
3	itrf2005	geodesy	**orbit determination**	67	**doris**	25
664	mission	precise orbit	system	1717	astrometry	778
4739	model	determination	**precise orbit determination**	15	gaia	28
204	network	polar motion	**model**	4739	**model**	4739
67	orbit determination	**orbit determination**	geodesy	13	precession	128
15	precise orbit determination	thermospheric model	network	204	orbit	188
129	pressure	grace	polar motion	11	**topex/poseidon**	22
302	satellite	envisat	**pressure**	129	methods:data analysis	1734
12	terrestrial reference frame	sea level	thermospheric model	11	radiation belts	25
22	topex/poseidon	gps	envisat	14	space	1024
		tracking	gps	94	service	23

The subject structure of a university – using Tongji University as a case study

Xinyue Xu [1], Pengfei He [2], ShiKang Ng [3] and Yuxian Liu [4]

[1] *xxy520@tongji.edu.cn*
Institute of Higher Education, Tongji University, Siping Road 1239, 200092 Shanghai (China)

[2] *ph232@tongji.edu.cn*
Office of Research Administration, Tongji University, Siping Road 1239, 200092 Shanghai (China)
[3] ngshikang@u.nus.edu
NUS Business School, National University of Singapore, Lower Kent Ridge Road 21, 119077 Singapore (Singapore)

[4] *yxliu@tongji.edu.cn*
Tongji University Library, Tongji University, Siping Road 1239, 200092 Shanghai (China)
School of History and Archives, Yunnan University, Cuihubeilu 2, 650091 Kunming (China)

Abstract

We use the subject categories of the journals that Tongji University's publications covered by Web of Science as the basic units to construct the university's subject structure in the framework of inter-category co-membership of journals. On the whole, Tongji University's subject structure is dominated by engineering. The four clusters, Environmental Engineering, Materials, Biomedicine, Engineering Management, forms the backbone of the structure. After slicing the 18 years of data into 3 six-year stages, subject structures were constructed for each stage. From the evolution of the structure, we can see that the biomedicine is developed rapidly; inducing management and other social sciences and fundamental sciences are clustered differentially as well as integratedly into these four clusters. Environmental Engineering is relatively stable in three stages. Engineering Management is more inclined to combine with science and engineering after three stages of development. Materials presents a complex and unstable situation because the diversity that materials are drawn from and its relationships with the fundamental sciences such as physics and chemistry are different. The characteristics consist with the actual development of the university. The construction and analysis of subject structure can provide a powerful reference for discipline-related planning of universities.

Introduction

Over the development of subjects' independence and integration, complex relationships between subjects were formed. Each subject is linked to another by these relationships to form a certain structure. This is the subject structure of science (Zhao, 1986). Efforts to outline a map of science using citation between publications can be traced back to the work by Garfield (1964) and Price (1965). When delineating a map of science, the fundamental elements on which it is based are called basic units. Early maps of science used academic publications as the basic unit. Boyack, et al. (2005) extended the basic unit to journals and constructed a backbone of science using inter-category citation. Leydesdorff & Rafols (2009) then used the subject categories and research areas in the Web of Science (WoS) as the basic unit, using principal component analysis on the citation data collected from the Science Citation Index-Expanded (SCIE) to construct the global map of science. Since Boyack, et al. (2005) used journal as a basic unit on the grounds that the topics published in the same journal were related to each other, Liu (2018) inferred that the interrelation of these topics in a journal also establishes a correlation between the subject categories the journal is assigned. She termed this relationship an inter-category co-membership of journals. While the relationships among citations are prone to changes, the relationship of subject categories based on the journals' assignment to the subject categories is a relatively stable. Liu (2018) illustrated that a global backbone of science constructed according to these newly defined relationships provides a more stable and reliable reference framework for disciplinary planning.

A global backbone of science can help us understand the logic and structure of knowledge systematically. Based on the logic and structure of science and technology (S&T) and its relationship, layout of the research and development (R&D) and educational systems can be arranged reasonably and all the disciplines and fields of S&T can be organized into a logical landscape. Such an organization will promote transdisciplinarity and the integration of different disciplines, greatly enhance the efficiency of scientific research, and accelerate the progress of S&T (Li, 2016).

Different universities in various countries and regions have formed their own unique subject structures due to the different research focus. Generally, publication is the main output of research. Using a university's publications to construct the subject structure of the university can understand how the subjects were related to each other. This provides a reference for tracking the development of the subject, for finding areas for future collaboration and for adjusting the subject structure according to the goal of the development.

The Chinese government launched the "Double First-Class" initiative in 2015. The aim of the initiative is to ultimately build a number of world-class universities and first-class disciplines by the end of 2050, in an effort to make China an international higher education power (People's Daily Online, 2017). The initiative regarded that the construction of first-class universities should be based on first-class disciplines. A university should grasp the law that steers the development of disciplines, and clarify the relation of the disciplines so that a harmonious and sustainable disciplinary system can be established (Ministry of Education, Ministry of Finance, National Development and Reform Commission, 2018).

The selected list of universities participating in the plan was released in September 2017 (Ministry of Education, et al. 2017). And then each university on the list issued its own construction strategy to become a first class university. However, these strategies are mostly based on qualitative analysis on the subject characteristics of the university. If we can depict the subject structure quantitatively, we can analyse the characteristics of subject structure more accurately and find a path between the subjects. Based on analysis, we can make a more powerful strategy to construct a world-class university.

Tongji University is one of the selected world-class universities located in Shanghai. It is a comprehensive research university which offers a wide range of programs in sciences, engineering, medicine, arts, law, economics and management. Using Tongji University as a case study, this study aims to quantitatively depict and analyse the subject structure of the university. We sketch the subject structure of Tongji University, constructed based on the 2001-2018 publications of Tongji University, which was then used to analyse the overall subject structure of Tongji University and relationships among various subject categories. After slicing the 18 years of data into 3 six-year stages, subject structure maps were constructed for each stage. Based on these maps, we will analyse the development of the subject structure, and the changes in the subject connections.

Data

The data used in this paper is from the Web of Science (WoS). By the end of October 2018, WoS covered 11,654 academic journals. These journals fall into 242 subject categories (SCs). Among them, there are 182 subject categories belonging to SCIE, containing 8,996 journals; there are 60 subject categories belonging to SSCI, containing 3,303 journals. At the same time, there are 15 subject categories belonging to both SCIE and SSCI. Overall, more than 5,261 journals belong to two or more subject categories. Among them, 3,789 journals belong to two different subject categories, 1117 journals belong to three different subject categories, 303 journals belong to four different subject categories, 43 journals belong to five different subject categories and 9 journals belong to six different subject category categories.

We conducted a search in the WoS and download all publications from Tongji University from 2001 to 2018 by setting the organization name and time period. Table 1 lists the number of publications, the number of journals involved, and the number of categories they belong to. Over these 18 years, Tongji University have 84,171 publications covered in WoS. These publications were published in 9,334 journals and belong to 225 subject categories.

Table 1. Statistics of publications from Tongji University from 2001 to 2018

Stage	*Number of publications*	*Number of journals*	*Number of SCs*
2001-2006	12188	1871	158
2007-2012	29342	4634	205
2013-2018	42641	6501	221
2001-2018（total）	84171	9334	225

Result

Based on the publications of Tongji University from 2001 to 2018, we used VOSviewer software to construct Tongji's subject structure. The software uses the distance between two items reflects the strength of the relation between the items (van Eck & Waltman, 2010). The software provides clustering technology. The number of clusters can be determined according to different settings of parameters such as resolution, random starts, and iteration. Leydesdorff et al. (2013) grouped the subject categories into 4 clusters, 6 clusters, and 19 clusters for user reference. After many trials, we have chosen to cluster the subject categories into six clusters to facilitate our analysis. We sliced data into six-year periods and constructed stage subject structures (Figures 1, 2, and 3) and the overall subject structure of Tongji University (Figure 4). The subject categories in WoS were used as the basic units and similarity between the subjects were calculated by the number of common journals they have.

Tongji University's subject structure constructed from its publications

Subject Structure and the Changes over time

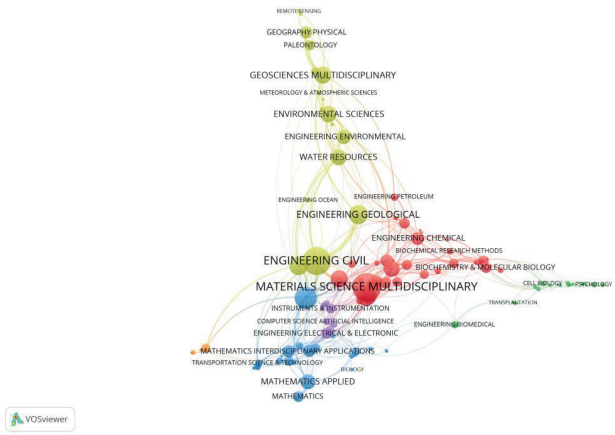


Figure 1. Tongji University's subject structure (2001-2006)

Table 2. Clustering of SCs of Tongji by VOSviewer (2001-2006)

Number	*Color*	*Cluster Name*	*Number of Categories*
1	Red	Materials	42
2	Green	Biomedicine	30
3	Blue	Engineering Management	29
4	Yellow	Environmental Engineering	21
5	Purple	Physics	13
6	Orange	Acoustic-radio pathology diagnosis technology	4

The subject structure for the period of 2001 to 2006 is shown in Figure 1. Table 2 shows the names of the six clusters. In this stage, the boundaries of the clusters are clear, and the positions of subject category from the same cluster are concentrated. The overall structure is roughly triangular in space, with Engineering Civil and Materials Science Multidisciplinary located at the centre of the structure.

The medical subjects are located on the left and right sides separately. Acoustics, Audiology & Speech-Language Pathology, Gastroenterology & Hepatology, Radiology Nuclear Medicine & Medical Imaging are clustered together. This cluster locates near the cluster of Engineering Management. It does not link to the other biomedical subjects directly, but via the subject of Acoustic, links to the clusters of physics, Environmental Engineering and Engineering Management. It seems that during this period that Tongji scholars were trying to understand the mechanism of the pathology diagnosis technology based on the image which is detect by the radio and acoustic wave. We call this cluster the Acoustic-radio pathology diagnosis technology.

The cluster of Engineering Management consists of 29 subject categories. The subjects involve STEM and Social sciences. Some are fundamental sciences such as physics and mathematics; some are applied sciences such as Engineering Industrial and Transportation Science & Technology. In the centre of the cluster is the subject of Operations Research & Management Science, which bonds the other subjects in this cluster together. Around the subject of Operations Research & Management Science are Engineering Industrial, Engineering Manufacturing and Mathematics Interdisciplinary Applications. The computer-related subjects are in the right side, the subjects of Social Sciences are in the left side. Mechanics is located in the top of this cluster, and Mathematics is located in the bottom of this cluster. The subjects of Physics are between the centre and Social Science. We use the main subjects in the centre of the cluster to name this cluster "Engineering Management". It is actually a strong subject in Tongji University.

Psychology, located at the rightmost within the cluster of Biomedicine, first connects to the subjects of Neuroscience and Cell Biology within the same cluster, then connects to the subject of Biotechnology in the Materials cluster, and further extends from Biotechnology to Biochemistry. There are two directions from Biotechnology: up through the subject of Engineering Chemical, connected to the cluster of Environmental Engineering which is more closed related to Chemistry; down through Chemistry Physical and Polymer Science, connected to the Cluster of Materials which is more closely related to physics, and then connected to its fundamental clusters: first to the cluster of Physics and then to the cluster of Engineering Management through the subject of Mathematics Interdisciplinary Applications. The logic that connects the clusters is very clear.

From the above analysis, we can see that: (1) The cluster of Materials, the cluster Engineering Management and cluster of Physics have obvious transitional characteristics. They are located in the centre of the structure. (2) The subjects at the top and right apex of the triangle structure are all concentrated in biomedicine and environmental engineering. (3) The subjects in the left

apex cannot be clustered normally, but it reflects the development characteristic of Tongji University when Biomedicine began to develop.

Figure 2 and Table 3 show subject structure of Tongji University from 2007 to 2012. Compared with the first stage, the boundaries of each cluster are not very clear. At this stage, the subject development of Tongji University presents a turbulent situation of differentiation and integration, and it is a transitional stage in which the links between major clusters are strengthened.

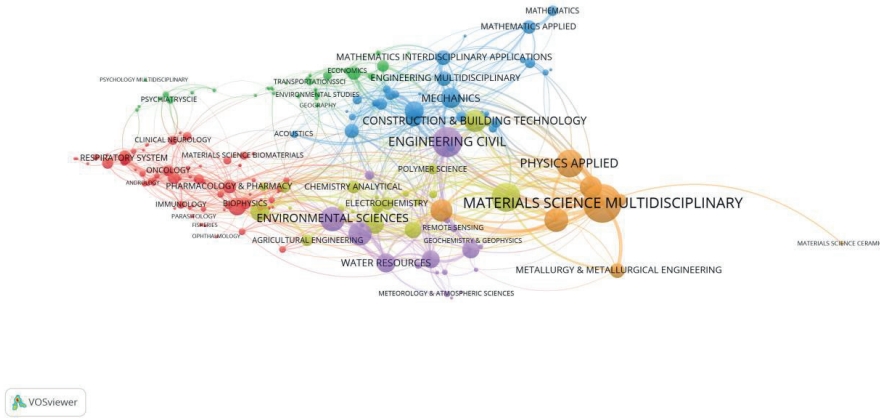


Figure 2. Tongji University's subject structure (2007-2012)

Table 3. Clustering of SCs of Tongji by VOSviewer (2007-2012)

Number	*Color*	*Cluster Name*	*Number of Categories*
1	Red	Biomedicine	74
2	Green	Engineering Management	38
3	Blue	Mechanical Engineering	32
4	Yellow	Chemical Engineering	23
5	Purple	Environmental Engineering	21
6	Orange	Materials	8

The subjects of Acoustics, Audiology & Speech-Language Pathology in the cluster of Acoustic-radio pathology diagnosis technology in the previous stage are clustered in the cluster of Mechanical Engineering, together with the other subjects of physics in this stage. The subjects of Gastroenterology & Hepatology, Radiology Nuclear Medicine & Medical Imaging are clustered into the Biomedicine cluster and have merged with it.

The subjects of computer science, mathematics and physics leave the cluster of Engineering Management and enter the cluster of Mechanical Engineering. The subjects of Management, Planning & Development are grouped into Engineering Management. The subjects of psychology move from the cluster of Biomedicine to Engineering Management.

The clusters of Mechanical Engineering, Chemical Engineering, Environmental Engineering and Materials present a trend of differentiation and integrations of disciplines. The cluster of Mechanical Engineering has many subjects of mathematics and physics, which are the

fundamental supportive subjects of this cluster. The subject of Chemistry Multidisciplinary is in the cluster of Materials, but its position is closer to the cluster of Chemical Engineering and the cluster of Environmental Engineering than to its own cluster. The subject of Engineering Civil is remains at the centre of the entire structure. It is grouped in the cluster of Environmental Engineering, but its position is closer to the cluster of Mechanical Engineering and the cluster of Chemical Engineering than to its own cluster. The subject of Construction & Building Technology is clustered into the cluster of Chemical Engineering, but it is inside of the cluster of Mechanical Engineering, and has grown closer to the subject of Engineering Civil in the cluster of Environmental Engineering. The Chemical Engineering cluster spans the entire structure.

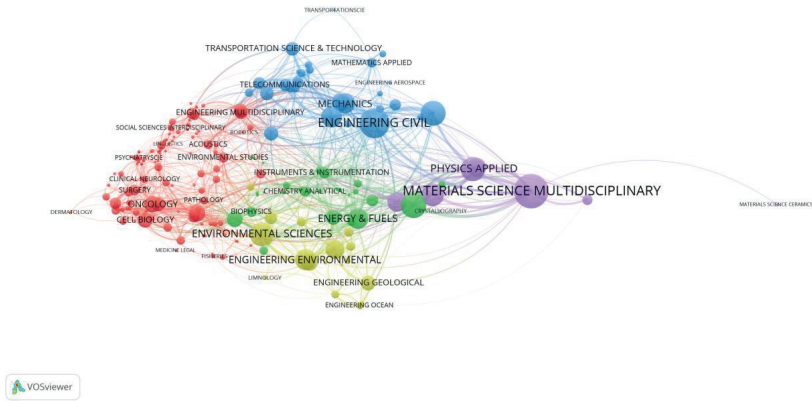


Figure 3. Tongji University's subject structure (2013-2018)

Table 4. Clustering of SCs of Tongji by VOSviewer (2013-2018)

Number	*Color*	*Cluster Name*	*Number of Categories*
1	Red	Biomedicine and Social Science	129
2	Green	Physics and Chemistry (inclus: Agriculture)	29
3	Blue	Engineering Management	27
4	Yellow	Environmental Engineering	20
5	Purple	Materials	9
6	Orange	Dermatology and Mycology	2

Figure 3 and Table 4 show subject structure of Tongji University from 2013 to 2018. The cluster boundaries have become clearer in this stage than previous stage.

Biomedicine and Social Science are clustered in one cluster. Some subjects of social sciences leave the engineering-related cluster and enter the cluster of Biomedicine, but these subjects are located in the between of the cluster of Biomedicine and the cluster of Engineering management. Neuroscience and Psychology have brought the subjects of social sciences and biomedicine together. In 2016, the Department of Psychology of Tongji University was formally established, with clinical psychology and psychological philosophy as the highlight of the discipline. With this, the connection between Biomedicine and Social Science was established. It seems the Biomedicine-related subjects developed so strong in this stage that it

attracted Social Sciences to it and make Social Sciences leave engineering-related clusters. The computer-related subjects enter the cluster of Engineering Management, making the cluster of Engineering Management incline to science and engineering. The subjects of social sciences in the cluster of Engineering Management reduced.

The cluster of Physics and Chemistry spans the structure and has the characteristic of obvious fusions. It is located in the centre of the structure. The subject of Chemistry Physical in this cluster is drawn closer to the cluster of Materials. The subject of Biotechnology & Applied Microbiology in this cluster is closer to the cluster of Biomedicine and Social Science. The subject of Agricultural Engineering in this cluster is closer to the cluster of Environmental Engineering.

The subjects of Dermatology and Mycology are not clustered into Biomedicine, but become a single cluster. Dermatology and Mycology were isolated or uncovered in the first stage. In the second stage, they were in the cluster of Biomedicine. In this stage, they become a independent cluster but still located beside the cluster of Biomedicine.

By comparing with the first two stages, the fundamental subjects such as physics, chemistry, and mathematics not only grouped into the cluster of Physics and Chemistry, but also dispersed into all the clusters. This shows that the links between fundamental sciences are combining into the related applied science.

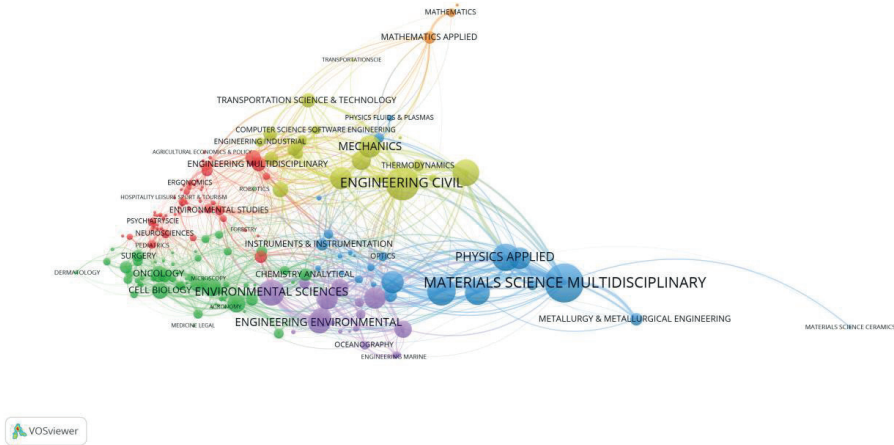


Figure 4. Tongji University's subject structure (2001-2018)

Table 5. Clustering of SCs of Tongji by VOSviewer (2001-2018)

Number	*Color*	*Cluster Name*	*Number of Categories*
1	Red	Social Science	75
2	Green	Biomedicine	74
3	Blue	Materials	25
4	Yellow	Engineering Management	23
5	Purple	Environmental Engineering	20
6	Orange	Mathematics	3

Figure 4 and Table 5 show the overall subject structure of Tongji University from 2001 to 2018. The overall subject structure has the following characteristics: (1) Comparing with the last stage,

Biomedicine and Social Science are separated into different clusters. (2). Three subjects, Mathematics, Mathematics Applied and Logic are grouped into a single cluster. The other mathematical subjects related to the concreted subjects, e.g. Social Sciences Mathematical Methods, are grouped into the relevant cluster. (3) Each cluster has clear boundaries, but there are also some subject categories which are closer to the other clusters in terms of the position e.g. the subject of Agricultural Engineering in the cluster of Biomedicine is closer to the cluster of Environmental Engineering than to its own cluster.

The Connections between subjects

The clusters are formed based on the connections between subjects. More connections between the subjects make them more easily clustered in one group. However, knowing how strength the links between the subjects help to know the mechanism that the subjects are connected and further understand how the interdisciplinary formed.

Table 6 shows the 10 pairs of subject categories with the strongest connections. There are 8 pairs are from the same clusters (intra-cluster connections). Two pairs are from the different clusters (inter-cluster connections). Materials Science Multidisciplinary in the cluster of Materials appears in top 10 pairs of subject categories 5 times, connecting with subject categories even from a different cluster, thus having inter-cluster connections with the cluster of Engineering management through subject categories such as Construction & Building Technology and Engineering Civil. All the top 10 pairs of connections occur in the clusters of Engineering management, Environmental Engineering and Materials.

Table 6. The 10 pairs of SCs with the strongest connections (2001-2018) （Number of Cluster refer to Table 5）

Subject Category 1	*Cluster*	*Subject Category 2*	*Cluster*	*Links*
Construction & Building Technology	4	Engineering Civil	4	3202
Environmental Sciences	5	Engineering Environmental	5	2040
Materials Science Multidisciplinary	3	Physics Applied	3	1700
Materials Science Multidisciplinary	3	Chemistry Physical	3	1680
Materials Science Multidisciplinary	3	Nanoscience & Nanotechnology	3	1614
Materials Science Multidisciplinary	3	Construction & Building Technology	4	1122
Engineering Geological	5	Geosciences Multidisciplinary	5	1024
Materials Science Multidisciplinary	3	Engineering Civil	4	964
Engineering Mechanical	4	Mechanics	4	954
Environmental Sciences	5	Water Resources	5	926

From further analysis of tables on the connections between subject categories, we can observe the following characteristics: (1) There are 5 isolated subjects, with no connections with the other subjects (Area Studies, Demography, Psychology Social, Substance Abuse SCIE, Substance Abuse SSCI). 8 subject categories only connected with one other subject category, which all occur within the same cluster. 271 pairs of subject categories only have two connections, among which, 62.7% occurs in the same cluster.

(2) Of the 5 isolated subjects, 4 belong to the Social Sciences. In the connections between Social Sciences and other subjects, 80% are intra-cluster connections, 20% are inter-cluster connections, and inter-cluster connections mainly concentrate within Biomedicine. It shows that the subjects of Social Science have a weak connectivity to other subject categories within Tongji University.

Table 7 shows a ranking of the top ten pairs of subject categories based on the number of links for each of the three stages. Through a comparison among the three stages, we can observe the following: (1) Compared to the first stage, the number of connections in the second and third phases has increased significantly. (2) Throughout the three stages, Engineering Civil, Construction & Building Technology, Environmental Sciences, Engineering Environmental, Materials Science Multidisciplinary, Physics Applied, etc. have always maintained strong relations with each other. They are all strong subjects of Tongji University.

Table 7. The top 10 pairs of SCs based on the links for three stages（Number of Clusters：2001-2006 refer to Figure 1；2007-2012 refer to Figure 2；2013-2018 refer to Figure 3）

Stage	*Subject Categories 1*	*Cluster*	*Subject Categories 2*	*Cluster*	*Links*
2001-2006	Engineering Civil	4	Engineering Geological	4	278
	Construction & Building Technology	1	Engineering Civil	4	112
	Materials Science Multidisciplinary	1	Physics Applied	1	112
	Construction & Building Technology	1	Materials Science Multidisciplinary	1	102
	Engineering Environmental	4	Environmental Sciences	4	102
	Mathematics Applied	3	Mathematics	3	102
	Engineering Civil	4	Engineering Mechanical	4	100
	Engineering Mechanical	4	Mechanics	3	96
	Geography Physical	3	Geosciences Multidisciplinary	4	86
	Materials Science Multidisciplinary	1	Metallurgy & Metallurgical Engineering	1	84
2007-2012	Environmental Sciences	5	Engineering Environmental	5	662
	Engineering Civil	5	Construction & Building Technology	4	624
	Materials Science Multidisciplinary	6	Nanoscience & Nanotechnology	6	458
	Materials Science Multidisciplinary	6	Physics Applied	6	454
	Materials Science Multidisciplinary	6	Chemistry Physical	4	426
	Materials Science Multidisciplinary	6	Metallurgy & Metallurgical Engineering	6	306
	Engineering Civil	5	Engineering Geological	5	298
	Materials Science Multidisciplinary	6	Physics Condensed Matter	6	256
	Physics Applied	6	Physics Condensed Matter	6	256
	Environmental Sciences	5	Water Resources	5	228
2013-2018	Construction & Building Technology	3	Engineering Civil	3	2466
	Engineering Environmental	4	Environmental Sciences	4	1276

Materials Science Multidisciplinary	5	Chemistry Physical	2	1204
Materials Science Multidisciplinary	5	Physics Applied	5	1134
Materials Science Multidisciplinary	5	Nanoscience & Nanotechnology	5	1108
Materials Science Multidisciplinary	5	Construction & Building Technology	3	824
Engineering Geological	4	Geosciences Multidisciplinary	4	794
Materials Science Multidisciplinary	5	Engineering Civil	3	760
Engineering Mechanical	3	Mechanics	3	700
Environmental Sciences	4	Water Resources	4	646

Conclusion

We use the subject categories of the journals that Tongji University's publications covered by Web of Science as the basic units to construct the university's subject structure in the framework of inter-category co-membership of journals. After slicing the 18 years of data into 3 six-year stages, we analyse the characteristics of the development of Tongji University's subject structure, in an attempt to provide a reference for the university's planning work on subject construction.

On the whole, Tongji University's subject structure is dominated by engineering. We group the subject categories covered by the publications of Tongji University into 6 clusters. Throughout the 18-year development process, the main line of the evolution in Tongji University's subject structure is Environmental Engineering, Materials, Biomedicine, and Engineering Management. Biomedicine is developed rapidly, management and other social sciences and fundamental sciences are clustered differentially as well as integratedly into the above four clusters in the different periods and the other two clusters are formed according to the characteristic of differentiation and integration at each stage.

The cluster of Environmental Engineering is relatively stable, and the subject categories included in three stages and overall stages are roughly the same, including Geography, Limnology, Oceanography and Remote Sensing, which is used to detect resources in the earth and ocean. However, the cluster of Materials presents a complex and unstable situation because the diversity that materials are drawn from and its relationships with the fundamental sciences such as physics and chemistry are different: one is that Materials-related subjects are grouped into different clusters, and the other is that the subject categories from the same cluster vary obviously. In the 2007-2012, 2013-2018 and 2001-2018 stages, the cluster of Materials mainly relate to ceramics, coatings & films, physics and chemistry, but in the 2001-2006, the cluster has the subjects related to biomedicine.

Biomedicine has developed rapidly and has sprung up during 2001-2018. In 2001-2006, Tongji University had a few subject categories related to biomedicine and these subject categories weren't concentrated. The biomedical subjects and acoustic-radio pathology diagnosis technology based on physics are divided into two clusters, which are distributed at the left and right ends of the structure. In 2007-2012, it has grown and expanded, not only has the number of subjects increased, but also some of the diagnosis technologies based on physics and health care were also gathered into this cluster. The cohesiveness of Biomedicine has become stronger. In 2013-2018, Biomedicine has brought most of the social sciences into own cluster through

Neuroscience and Psychology. By combining the three stages of data, the Biomedicine and Social Science are clustered into two clusters, but they are still adjacent.

Engineering Management is a discipline that integrating social sciences and engineering. In 2001-2006, this discipline includes social sciences such as Business Finance, Economics, Social Sciences Mathematical Methods, fundamental sciences such as Mathematics and Logic, and applied sciences such as Engineering Industrial, Engineering Manufacturing, Computer Science-related subjects. In 2007-2012, the psychological subjects enter the cluster of Engineering Management, computer science, mathematics, physics and other subjects leave this cluster and enter the cluster of Mechanical Engineering. In 2013-2018, the social sciences in the cluster of Engineering Management decreased, and the subjects related to computer science once again enter this cluster, making this cluster more bias towards science and engineering. The subject categories of this cluster at this stage are consistent with the subject categories of this cluster during 2001-2018.

The other two clusters are under the impact of medicine, and take shape because of the combination of these four clusters and basic subjects, for example, physical and acoustic diagnosis in the first stage. And Chemical Engineering and Mechanical Engineering in the second stage are formed by the combination of chemistry, physics and technology. In the third stage, the cluster of Physical Chemistry and Dermatology and Mycology are cluster specially.

Engineering Civil is a strong discipline of Tongji University. It is clustered in Environmental Engineering in the first, second stage and the whole stage, and take on a central position. Although it does not become a single cluster, the connections with other subjects are always the strongest, and they are mainly inter-cluster connection. Engineering Civil, Construction & Building Technology, and Environmental Sciences (all under science and engineering) are closely related to other subjects, showing a stronger development trend. The strength of links for the social sciences is improving slightly, although there is a gradual increase trend, the difference is still significant compared with science and engineering.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (No.81573023 and 71173154). We also thank Tongji University's Funding for attending the international conferences.

References

Boyack K W, Klavans R, Börner K.(2005). Mapping the backbone of science. *Scientometrics*, 64(3), 351-374.

Gafield E, Sher I H, Torpie R J. (1964). *The use of citation data in writing the history of science*. Philadelphia:Institute for Scientific Information Inc..

Jinghai Li (2016).Exploring the Logic and La ndscape of the Knowledge System:Multilevel Structures, Each Multiscaled with Complexity at the Mesoscale. *Engineering*, 2, 276-285.

Leydesdorff L. & Rafols I., (2009). A global map of science based on the ISI subject categories. *Journal of the American Society for Information Science and Technology*, 60(2), 348-362.

Loet Leydesdorff, Stephen Carley, Ismael Rafols. (2013). Global maps of science based on the new Web-of-Science categories. *Scientometrics*, 94, 589–593.

Liu YX. (2018). Constructing a global backbone of science based on inter-category co-membership of journals. *Journal of The China Society for Scientific and Technical Information,* 37(6), 580-589.

Ministry of Education, Ministry of Finance, National Development and Reform Commission. (2017). Ministry of Education, Ministry of Finance, National Development and Reform Commission released *a selected list of universities and colleges, which will participate in the country's construction plan of world-class universities and first-class disciplines*. : http://www.moe.gov.cn/srcsite/A22/moe_843/201709/t20170921_314942.html

Ministry of Education, Ministry of Finance, National Development and Reform Commission. (2018). Notice from Ministry of Education, Ministry of Finance, National Development and Reform

Commission announcing *Guidance on Speeding up the Construction of "Double First-Class" in institutions of higher learning*: http://www.moe.gov.cn/srcsite/A22/moe_843/201808/t20180823_345987.html
Price D J.(1965). Networks of scientific papers. *Science*, 149, 510-515.
van Eck, N.J. & Waltman, L. (2010). Software survey: VOSviewer, a computer program for bibliometric mapping, *Scientometrics*, 84, 523–538.
People's Daily Online. (2017). China to develop 42 world-class universities. People's Daily Online. 21 September 2017. http://en.people.cn/n3/2017/0921/c90000-9272101.html
School of Humanities, Tongji University, College overview, Organization, http://sal.tongji.edu.cn/index.php?classid=4515&showorg=66
Tongji University, Facts and figures, https://en.tongji.edu.cn/About_Tongji/Facts_and_Figures.htm
Web of Science Core Collection Help: Research Areas (Categories / Classification) http://images.webofknowledge.com//WOKRS531OR13/help/zh_CN/WOS/hp_research_areas_easca.html
Zhao HZ. (1986). On the science structure. *Academic Journal of Zhongzhou,* 1981(3), 59-65.

Persistent Problems for a Bibliometrics of Social Sciences and Humanities and How to Overcome Them

Jochen Gläser[1] and Jenny Oltersdorf[2]

[1] *Jochen.Glaeser@tu-berlin.de*
TU Berlin, Hardenbergstr.16-18, 10623 Berlin (Germany)

[2] *Jenny.Oltersdorf@tu-berlin.de*
TU Berlin, Hardenbergstr.16-18, 10623 Berlin (Germany)

Abstract
Evaluative and structural bibliometrics respond to their inability to analyse the social sciences and humanities with three strategies, namely (1) ignoring the problem and treating them with the methods developed for the sciences; (2) focussing on the problem that seems easiest to overcome, namely coverage by commercial databases; and (3) studying SSH publications with the aim to better understand SSH publication behaviour. In this paper, we argue that these responses express a principal limitation of bibliometrics, namely the artificial separation of research and publication practices it creates by focussing on the artefacts produced by publication practices. We use empirical material from previous studies to outline the research problem underlying the helplessness of bibliometrics, and suggest a research programme that could provide theoretical foundations for bibliometric analyses of the SSH.

Introduction

For almost two decades, the necessity of treating SSH differently has been widely acknowledged by bibliometricians, particularly in the light of the diversity of publication types and languages, which is linked to the limited coverage of the SSH literature by the big citation databases (Frost 1979, Garfield 1980, Stern 1983). The suggestion to abandon citation-based evaluations altogether and to use alternative approaches (Moed et al. 2002) has been ignored. Instead, bibliometric studies have responded to these problems in three different ways:

(1) Ignoring the problem
Quite surprisingly, bibliometricians keep publishing studies that analysed the SSH as if the problem did not exist. The simplest form of ignoring the problem is not mentioning it at all and conducting standard publication and citation analyses (Sternberg and Litzenberger 2005, Daniel 2006). Other authors notice that they base their analyses on a small proportion of the literature and provide arguments why the analysis should be conducted nevertheless, e.g. "because it is better to know at least something of a small portion of the output, than to have no insight in the impact of these papers at all" (Van Leeuwen 2006: 152) or because the datasets are "hopefully representative of the research published in the respective specialism" (Colavizza et al. 2019: 108).

(2) Attempting to extend the corpus
Bibliometricians also attempt to improve citation studies by including citations to non-source items (Butler and Visser 2006, Linmans 2010), to improve network analyses by matching non-source items (Kristensen 2018, Colavizza et al. 2019), and to improve evaluation studies by including library holdings of books (White et al. 2009, Linmans 2010, Zuccala and Cornacchia 2016). None of these attempts solves the fundamental problem described above.

(3) Studying publications with the aim to better understand SSH publication practices
The obvious peculiarities of the SSH literature have triggered many studies that attempt to identify and understand the publication and citation behaviour of SSH scholars. Many of these studies are based on manual collections of SSH literature and proceed by categorising publications and citations. Other studies include citation context analyses or surveys of researchers. We will discuss this literature more extensively in the following section.

These three types of responses are based on different interpretations of the problem faced by bibliometricians. The first response – ignoring the problem – expresses the belief that the limited and uneven coverage of the SSH literature does not distort analyses to an extent that invalidates findings. The second response expresses the belief that insufficient coverage is the root problem that needs to be solved. The third response is based on the belief that a better description of publication and citation practices creates the knowledge that is necessary to conduct valid bibliometric analyses of the SSH. It might also express the belief that this is as far as bibliometrics can go with its understanding of publications.

In this paper, we will argue that the third approach has turned up enough evidence to consider strategies (1) and (2) untenable, and strategy (3) insufficient. The evidence accumulated by descriptive studies of SSH publication practices and our own explorations of the link between research practices and publication practices in German art history suggest that without a theory of SSH knowledge production, we will be unable to understand the meaning and functions of publications and citations. Without such an understanding, neither the attempt by structural bibliometrics to identify knowledge structures and knowledge flows nor the attempt by evaluative bibliometrics to measure research performance are likely to be successful for SSH fields.

Therefore, the aim of our paper is to outline the research problem we face and a research programme that might successfully tackle the problem. We use empirical material from studies on research and publication practices in various SSH fields. We begin by briefly presenting findings of studies applying the third strategy, which are followed by some of our own findings on the (in)visibility of German art history to citation databases and the reasons why it is scarcely visible. From these empirical findings, we derive a research problem and a research programme.

Major properties of SSH publication practices

We begin our argument by presenting results of studies of SSH publication and referencing practices. These studies' findings clearly demonstrate that ignoring the differences between SSH and the sciences or extending the corpus by whatever sources are easily available cannot lead to valid results, thereby discrediting strategies (1) and (2). Studies of publication practices and of practices of referencing the literature have employed three main approaches:

- Samples of publications indexed by databases or purpose-built corpora of SSH literature or references in SSH publications have been categorised according to properties of publications. These studies primarily identified types of publications or referenced materials, their age characteristics, the frequency at which they were used, and the languages in which they were published (Broadus 1971, Garfield 1980, Heinzkill 1980, Stern 1983, Ardanuy et al. 2009, Hammarfelt 2012, Sivertsen 2016).
- The use of the literature by those citing it has been studied by citation context analyses, which categorised the importance of cited references for the citing text, the sentiment of the citing author, the distribution of citations across different sections of a publication or the functions of the cited literature in the citing publication (Frost 1979, Cozzens 1985, Hyland 1999).

- Motivations for citing particular publications in specific ways have been explored by standardised surveys and interviews (Brooks 1985, Shadish et al. 1995, Harwood 2009).

The studies applying these approaches revealed distinct properties of SSH publications (for overviews, see Hicks 1999, 2004, Nederhof 2006, Van Leeuwen 2006, 2013, Hammarfelt 2016). Their main results were that SSH
- use a variety of publication types, with book publications being the most frequently used type;
- often address nationally or regionally specific topics;
- publish in national languages;
- include primary sources in their reference lists;
- use older references than the sciences, which is only partly due to the referencing of primary sources; and
- are covered by the major citation databases (Web of Science and Scopus) only to a very limited extent.

There is also some more specific evidence for differences between citation practices of SSH and the sciences. SSH have been reported to use more negative citations (but still rather few, Brooks 1985, Hyland 1999, Bornmann and Daniel 2008: 54-56, Hellqvist 2010: 313). Hargens (2000: 858-860) found that authors from the social sciences (but not from the humanities field) use "orienting reference lists" (lists of papers as examples of a general perspective, methodological approach, or topic) more often than the sciences.

Generally, citation context analyses suffer from coarse categorisation schemes that are of little theoretical value (for an extensive report of such schemes, see Bornmann and Daniel 2008). Results obtained by citation context analyses remain descriptive and offer little more than repeated confirmation of the observation that publications are cited for a variety of reasons.

Investigations of citer motives are similarly descriptive. This has partly to do with the predominant use of lists of predefined categories in questionnaires and interviews. A study based on open interviews and a bottom-up categorisation led to a more sophisticated account of citer motives and to the observation that more than half of the citations had more than one function. This study, too remained descriptive (Harwood 2009). Apparently, there is no suitable theoretical framing for explanatory studies that link research practices and communication contexts underlying a publication to citer motivation. The old and meanwhile stale debate about a 'normative' versus 'constructivist' theory of citation (see e.g. Bornmann and Daniel 2008) reduces the issue to an extent that makes a nuanced analysis of functions of citations impossible.

Hellqvist (2010) attempted to build a theoretical account of citation practices by invoking Whitley's (Whitley 2000 [1984]) categorisation of fields according to their functional and strategic mutual dependency.[i] Whitley described the humanities as fragmented adhocracies with low control over standards and concepts, and as linked to a plurality of diverse audiences. Hellqvist links these properties to greater intellectual freedom, which he links to the more extensive and diverse citation practices of the humanities (Hellqvist 2010: 313-314). A second property Hellqvist considers is the individualistic nature of humanities research and the specific notions of originality identified by Guetzkow et al. (2004). Unfortunately, the link between his theoretical account and empirical findings, although suggestive, is not strong enough to support an explanation. Explaining empirical evidence with Whitley's theoretical categorisation of the humanities would have required operationalising the framework for a comparative investigation of citation practices.

This very brief account of the state of the art on publication and citation practices of SSH fields enables two conclusions. First, the distinctiveness of these practices makes it likely that the functions of publications for the author citing them are at least partly different from those in the sciences. These differences may be due to specific SSH research practices including specific links between research and publication practices. If this is the case, any information derived from bibliometric analyses, be they structural or evaluative, cannot be interpreted without systematic knowledge of these specifics. Second, systematic knowledge of the specifics of SSH knowledge production must be more concrete than the general theoretical accounts currently available, which reduce the description of SSH research to few variables (Whitley 2000 [1984], Becher and Trowler 2001). In the following, we use an empirical investigation of German art history to illustrate these points.

Methods

The main source of empirical information is the PhD project by Jenny Oltersdorf (Oltersdorf 2013) on German art history. We present data on the coverage of German art history by the Web of Knowledge (WoK), interview data on German art historians' research and publication practices, and data from documents describing research practices of art historians. The interviews covered three main topics, namely the interviewee's research practices (the ways in which they produce new knowledge), the functions of the different types of publications identified in the analysis of publications, and the ways in which art historians acquire and assess knowledge offered by their colleagues.

Analysis of coverage: Of the 127 professors of art history who could be identified through an internet search of all universities in 2010, 101 (79.5%) had lists of publications on their websites, which were categorised according to publication type and publication language. Of these, all publications between 2000 and 2009 were compared to publications by the 101 authors indexed in the Web of Science (SCI-Expanded, SSCI, A&HCI, BKCI, Search conducted in August and September 2012). The search was conducted as author search. Titles of retrieved publications were compared to authors' publication lists. A separate search based on the ISSNs of journals and publication year was conducted for all publications that could be expected to be found because the journal was indexed but weren't found with the author search. The same approach was used for a search in Scopus. Since Scopus returned fewer results, we report only the results for the WoS search. To better understand the publication practices of German art historians, five interviews with professors of the history of arts were conducted in 2012.

Findings: The invisibility of German art history to bibliometrics and some of its reasons

Invisibility of German Art History to International Databases due to Publication Types and Publication Languages

Table 1 lists the categories of publications from personal lists of German art historians, the numbers of publications found in each category, and the numbers of publications in each category returned by the search of the Web of Knowledge. It is no exaggeration to state that German art history is effectively invisible to the Web of Knowledge.

Table 1. Coverage of publications of 101 German professors of art history by the Web of Knowledge (including Book Citation Index)

Category	*Number of publications found in lists*	*Number of publications found in WoS*	*% Coverage*
Monographs	180	0	0
Book chapters	1540	7	0.5
Journal articles	352	70	19.9
Book reviews	301	31	10.3
Encyclopaedia entries	93	0	0
Entries in art catalogues	390	0	0
Newspaper articles	100	0	0
Total	2956	108	3.7

These results are plausible given the publication languages. Of the 2956 publications, 90% were written in German, 5% in English, 3% in Italian and 2% in French. This raises the question as to how international communication is realised in art history, which can be expected to be an international field. A first hint is provided by the following statement of the *Comité International d'Histoire de l'Art*/International Committee for Art History (CIHA)[ii]

„The language used for CIHA activities should reflect and promote national and international diversity, exchange and understanding." (CIHA 2016)

Official languages in CIHA are German, English, Spanish, French and Italian (Dufrene 2007). The Journal of another organisation, *the International Association of Research Institutes in the History of Art* (RIHA) welcomes contributions in any of these five languages. It even accepts articles in other languages but encourages parallel publication in one of the CIHA languages.

„Articles are published in one of the five languages authorized by the Comité International d'Histoire de l'Art (CIHA), i.e. in English, French, German, Italian, or Spanish ("CIHA languages"). However, with regard to manuscripts written in a non-CIHA language, RIHA Journal encourages parallel publications in both the original language and in translation (e.g., Polish-English). Translations have to be arranged for by the author (or guest-editor of a special issue), and are required to be carried out by a language professional." (Journal of the International Association of Research Institute in the History of Arts 2019)

These statements are confirmed by the range of languages authors in art history appear to read. Table 2 lists the languages of references in the articles published in the November 2018 issue of the journal Art History.

Reasons for invisibility: Publication practices

The few interviews on research and publication practices provide already some indications why researchers publish in formats and languages that make their output invisible to citation databases. Commenting on publication types that communicate new knowledge, interviewees characterised monographs as the main carrier of new knowledge in art history. This is why book reviews are also important: They are an efficient means of telling the community what new knowledge can be found in which book. A second, minor source of new knowledge are entries in art catalogues. These entries may contain pieces of new knowledge.

Table 2. Publication languages found in endnotes of articles in Issue 5 of Volume 41 of Art history (November 2018)

Country of residence of the author	*Languages of references*	*Title*
United States	English, Dutch, French, German, Italian, Portuguese, Spanish	Common Threads: Cloth, Colour, and the Slave Trade in early Modern Kongo and Angola
United States	English, French, German, Italian, Latin, Russian	Phantoms of Emptiness: The Space of the Imaginary in Latin Medival Art
United States	English, French, Spanish	Luc Chessex, Robert Frank, and the Representation of Labour in the Magazine Cuba Internacional, 1968-71
United States	English, French, Italian,	Giorgio de Chirico's 'Jewish Hour': Metaphysical Painting in Ferrara, 1915-18
United States	English, Russian	Vasily Surikov and the Russian Point of View
United Kingdom	English, French	Girodet's Galvanized Bodies

Interviewees considered book chapters as records of conferences. They rarely contain new knowledge and are produced mainly because early career researchers need a record of publications. Journal articles were considered as secondary because they contain no new knowledge. English-language articles were given to professional translators and 'disowned' by art historians, i.e. not really considered as their publications. The latter attitude is also expressed by art historians reporting journal articles not at all or incompletely in their publication lists.

The main reason for using the German language given by interviewees was precision. Interviewees reported that they feel they cannot express their arguments in a foreign language with the necessary precision. We will come back to this issue below.

Finally, the interviewed German art historians did not really see the advantage of having peer reviews. Since part of their research consists in creating unique perspectives on works of art, there is little peer review can do to improve their arguments.

Causes of publication practices: Research practices

Methods in art history are manifold and the field is far from having a canon of authoritative methods. The method used by a researcher and the form in which it is used depend on the research question and the evidence a researcher wants to bring to bear on the question. Since research questions, decisions on what counts as evidence, and methods for collecting and interpreting evidence strongly depend on the personal perspectives of researchers (Gläser et al. 2010, Gläser et al. 2018), research processes and outcomes are far less standardised than in the sciences.

Art history methods are roughly divided into the three categories: object determination, object preservation and object interpretation. One approach to object interpretation is the iconographic-iconological analysis. Erwin Panofsky described the method with a three-step model of pre-iconographical description, iconographical analysis and iconological analysis. The pre-iconographical description of an image is the process of taking inventory of the physical

aspects of an image to describe the fundamental visual elements in the two-dimensional design. In iconographical analysis, motifs are assigned to specific themes or concepts based on the pre-iconographical description and the reconstruction of "what the authors of these representations had read or otherwise knew" (Panofsky 1955: 36) The real act of interpretation is iconology, which reconstructs the "intrinsic meaning or content" (Panofsky 1955: 40). Iconology implies an interpretive achievement that synthesises iconography, historical, literary and psychological methods. Interpretation is, to a large extent, a subjective process (Panofsky 1955: 38).

Iconography and iconology require the art historian
- to address the fundamental hermeneutic problem of translating images into language (Boehm 1978: 447);
- to interpret images, often by using methods and knowledge from several disciplines;
- to use all accessible sources, which usually includes primary sources and secondary sources from several disciplines; and
- to confront existing scholarship.[iii]

This brief description of iconography and iconology highlights some consequences for publication practices. Translating images into language, interpreting images, and using all accessible sources for recovering the former meaning of a work of art are likely to require a long argument that does not fit a journal article or book chapter. The hermeneutic problems involved also make the language into which images are translated a tool for the analysis. The necessary precision of wielding this tool is best achieved with the researcher's mother tongue, which requires a publication in that language.[iv] The translation of images in language also entails that the analysis is not separable from the process of writing the publication. Using all available sources requires the command of several languages, which contributes to explaining the multilingual form of internationality that is characteristic of art history.

Discussion: The research problem faced by bibliometrics

While the empirical evidence we provided is far too sketchy to draw conclusions about the research and publication practices of SSH scholars, it makes it possible to state a research problem and to outline a research programme that can provide the knowledge necessary for bibliometric analyses of SSH to become meaningful.

Both the problem and the programme start by challenging an implicit premise bibliometrics and the sociology of science appear to agree upon, namely the assumption that research in the sciences, social sciences and humanities is conducted by individuals or groups who produce new knowledge and then communicate it to their colleagues via publications (and, to a lesser extent, by conference presentations, informal communication and so on). This premise enabled the separation of bibliometrics (studying the distribution and accumulation of published knowledge in scientific communities) from the sociology of science (studying the production of knowledge by individuals and groups).

In spite of its apparent plausibility, however, this premise could be wrong. Since any single knowledge claim receives its meaning from its embeddedness in a system of pre-existing knowledge claims, the notion of knowledge claims as separate products seems difficult to maintain. Instead, the production of knowledge in the sciences, social sciences and humanities is better understood as the collective undertaking of scientific communities, whose members jointly advance a shared body of knowledge (Gläser 2006, 2007, 2019). The important consequence of this shift of perspective is that communicating knowledge must be considered

as part of the production process and cannot be understood without considering its embeddedness in this process.

This is why the distinctiveness of SSH publications is not merely a technical problem for bibliometrics that can be solved by increased coverage of the literature. Bibliometrics is based on an implicit model of research that has been described by Collins (1994) as "rapid discovery science" – research driven by methodological development and based on the direct use of colleagues' findings. This research is dominated by one type of contribution, which is communicated by one type of publication. Since the corresponding model for the SSH does not yet exist, and since it is not even known yet whether one model would suffice, a mere technical adaptation of bibliometric approaches has small chances of success. In the words of the three strategies described in the introduction: Ignoring the problem does not lead to valid studies, extending the corpus without understanding SSH research still leaves bibliometricians with a sample of unknown representativeness, and describing publication and citation practices accumulates idiosyncratic snapshots unless the underlying regularities in research practices are known.

Research problem

As a collective production process, the production of knowledge is a two-level phenomenon of individual (or group) activities being integrated in a community-level process. Since publishing is a central mechanism linking the micro- to the macro-level, the research problem confronting the bibliometrics of SSH publications (or, for that matter, confronting bibliometrics in general) can be phrased as the following question: *How (by what mechanisms and with what effects) do publications link the micro-to the macro-level of knowledge production in SSH fields?* This questions points to three theoretical and empirical tasks. First, the productive function of publications consists in the knowledge flows they organise within scientific communities. Second, identifying these knowledge flows makes possible to study the interdependence between researchers, which is one or of the major social-structural property of scientific communities (Whitley 2000 [1984], Gläser et al. 2018). Third, these functions are field-specific, i.e. the productive functions of publications depend on epistemic conditions of action and social structures that are particular to fields of research in SSH.

Research programme

The empirical research programme for address this problem is partly specific to the SSH due to the diversity of their publication channels. Specifying the research problem described above as an empirical research programme for SSH publications leads to the following questions:

(1) What kinds of arguments are transmitted through which publication channels?
The empirical evidence pointed to different types of SSH publications containing different kinds of arguments and thus different types of new knowledge. In many SSH fields, books appear to contain the main arguments, while book reviews and journal articles may function as pointers to these arguments and the topics to which they apply. Since conferences and edited books often experiment with new conceptual frames for existing knowledge, book chapters may primarily contain knowledge that has been published elsewhere but is re-contextualised according to the frame suggested by a conference or theme for a book. New empirical knowledge or minor arguments might be published in any format.

(2) How are different kinds of knowledge used in individual-level production processes?
The major distinction between conceptual, methodological and empirical knowledge known from the sciences may be supplemented by further differentiations in SSH fields. This raises

the questions as to what types of knowledge exist in different SSH fields, in which types of publications they are transmitted, and how SSH scholars make use of these different types of knowledge in their own work. Answering these questions leads to a deeper understanding of the interdependence between SSH scholars that is mediated by publications.

(3) How are the individual-level research and publication practices interlinked?
The classical relationship of publications describing results of research processes ex post, which is ascribed to the sciences, does not hold for many SSH fields. If research is conceptual and language a major tool of analysis, the process of writing a publication is likely to be part of the analytical work, i.e. results emerge while scholars write their publication. This implies that the decision about a publication format influences the analysis, and that external control of publication formats (e.g. through evaluations) may have a much stronger influence on actual research practices in SSH fields than in the sciences, where publication channels are selected based on results.

These three questions suggest the combination of quantitative and qualitative methods of science studies, only some of which are genuine bibliometric methods:
- Citation relations such as direct citations or bibliographic coupling and text-based relations like co-word occurrence or lexical coupling can be used to model and analyse multipartite networks of different types of publications.
- Citation context analyses using content-based and function-based typologies of citation context can be used to identify the knowledge claims from cited publications used by citing authors and the ways in which these knowledge claims are used. Together with the network analysis described above, this enables the identification of knowledge flows and of the publication types utilised for the transmission of specific knowledge.
- Interviews with SSH scholars and observations of research processes can be used to identify the research practices that lead to specific knowledge and its communication in different types of publications. In connection with the other methods (and only in connection with them!) these qualitative methods can identify the mechanisms linking the micro to the macro level of SSH knowledge production.

The methods need to be applied in field-comparative studies. The diversity of SSH research and publication practices (see above, state of the art) makes it unlikely that functions of specific publication channels, types of knowledge claims produced, or mechanisms linking the micro level to the macro level apply equally to all fields. Field-comparative studies are therefore essential to establish causality.

Comparisons of national scientific communities are equally important. For example, the argument about the link between research methods and the major role of books in the national language, which has been made for German art history, would suggest that the variation between countries of publication types and proportions of English-language publications should be limited. However, such variation exists. This is why comparisons between countries in which humanities have a high proportion of English-language journal articles and countries like Germany are important to verify or specify that argument.

Conclusions

With this paper, we wanted to point out that bibliometric analyses, be they structural or evaluative, are forced to dramatically reduce the diversity of publication and referencing practices and to assign citations very limited meanings (Bornmann and Daniel 2008). Explorations of the underlying diversity have established the problem but could not go beyond

descriptive accounts because there is no suitable theoretical framework. We argued for addressing the underlying research problem and outlined a research programme that could help addressing it.

We would like to draw two conclusions. First, from our consideration of foundations for understanding publication practices in SSH fields follows that bibliometrics is unable to build its own theoretical foundations because it is a methodologically oriented field focused on a particular part of collective research processes and its theoretical foundations must embed the part of research processes it is concerned with into a theory of the whole research process. Thus, progress in building theoretical foundations of bibliometrics crucially depends on the collaboration with other fields of science studies.

Second, the argument made for SSH fields holds for the sciences and engineering, too. Although publication practices are undoubtedly more homogenous in these fields, the meanings and functions of citations appear to be much more diverse than the implicit model of bibliometrics and citation classifications building on that model suggest. While evaluative bibliometrics may resort to the statistical argument that citation counts are correlated to impact for sufficiently large collections of publications (Van Raan 1996, Gläser and Laudel 2007, Bornmann and Daniel 2008), any more differentiated analysis of impact and all structural investigations that want to reconstruct knowledge structures in publication collections need to address the problem of diversity of citations. This is a challenge to bibliometrics – but in the sense of a difficult research programme rather than a threat to its validity.

Acknowledgments

This research was supported by the German Federal Ministry of Education and Research (Grant 01PU17022).

References

Ardanuy, J., Urbano, C. & Quintana, L. (2009). A citation analysis of Catalan literary studies (1974-2003): Towards a bibliometrics of humanities studies in minority languages. *Scientometrics*, 81(2), 347.

Becher, T. & Trowler, P. R. (2001). *Academic Tribes and Territories: Intellectual enquiry and the culture of disciplines.* Buckingham: Open University Press.

Belting, H. (1995). *Das Ende der Kunstgeschichte. Eine Revision nach zehn Jahren.* München: C.H. Beck.

Boehm, G. (1978). Zu einer Hermeneutik des Bildes. In H.-G. Gadamer and G. Boehm (Eds.), *Seminar: Die Hermeneutik und die Wissenschaften* (pp. 444-471). Frankfurt a.M.: Suhrkamp.

Bornmann, L. & Daniel, H. D. (2008). What do citation counts measure? A review of studies on citing behavior. *Journal of Documentation*, 64(1), 45-80.

Broadus, R. N. (1971). The literature of the social sciences: a survey of citation studies. *International Social Science Journal*, 23(2), 236-243.

Brooks, T. A. (1985). Private Acts and Public Objects: An Investigation of Citer Motivations. *Journal of the American Society for Information Science*, 36(4), 223-229.

Butler, L. & Visser, M. S. (2006). Extending citation analysis to non-source items. *Scientometrics*, 66(2), 327-343.

Colavizza, G., Franssen, T. & van Leeuwen, T. (2019). An empirical investigation of the tribes and their territories: Are research specialisms rural and urban? *Journal of Informetrics*,13(1), 105-117.

Collins, H. M. (1994). Scene from Afar. *Social Studies of Science*, 24, 369-389.

Comité International d'Histoire de l'Art (2016). *CIHA Statutes. Official languages of the CIHA*. Retrieved January 25, 2019 from http://www.ciha.org/statutes.

Cozzens, S. E. (1985). Comparing the Sciences: Citation Context Analysis of Papers from Neuropharmacology and the Sociology of Science. *Social Studies of Science*, 15(1), 127-153.
Daniel, H.-D. (2006). The referee's comments on Gläser's criticism. *Scientometrics*, 67(2), 331-333.
Dufrene, T. (2007). *A Short History of CIHA*. Retrieved January 25, 2019 from http://www.ciha.org/sites/default/files/files/Short_History_of_CIHA.pdf.
Frost, C. O. (1979). The Use of Citations in Literary Research: A preliminary classification of citation functions. *Library Quarterly*, 49(4), 399-414.
Garfield, E. (1980). Is Information Retrieval in the Arts and Humanities Inherently Different from that in Science? The Effect that ISI'S Citation Index for the Arts and Humanities is Expected to Have on Future Scholarship. *Library Quarterly*, 50(1), 40-57.
Gläser, J. (2006). *Wissenschaftliche Produktionsgemeinschaften: Die soziale Ordnung der Forschung.* Frankfurt a. M.: Campus.
Gläser, J. (2007). The social orders of research evaluation systems. In R. Whitley & J. Gläser., *The Changing Governance of the Sciences: The Advent of Research Evaluation Systems* (pp. 245-266). Dordrecht: Springer.
Gläser, J. (2019). How Can Governance Change Research Content? Linking Science Policy Studies to the Sociology of Science. In W. Canzler, S. Kuhlmann & D. Simon. (Eds.), *Handbook of Science and Public Policy*. Cheltenham: Edward Elgar.
Gläser, J., Lange, S., Laudel, G. & Schimank, U. (2010). The Limits of Universality: How field-specific epistemic conditions affect Authority Relations and their Consequences. In R. Whitley, J. Gläser and L. Engwall (Eds.), *Reconfiguring Knowledge Production: Changing authority relationships in the sciences and their consequences for intellectual innovation* (pp. 291-324). Oxford: Oxford University Press.
Gläser, J. & Laudel, G. (2007). The social construction of bibliometric evaluations. In R. Whitley & J. Gläser. (Eds.), *The Changing Governance of the Sciences: The Advent of Research Evaluation Systems* (pp. 101-123). Dordrecht: Springer.
Gläser, J., Laudel, G., Grieser, C. & Meyer, U. (2018). Scientific fields as epistemic regimes: new opportunities for comparative science studies. *The Technical University Technology Studies Working Papers*, TUTS-WP-3-2018, Berlin: Technical University Berlin.
Guetzkow, J., Lamont, M. & Mallard, G. (2004). What is originality in the Humanities and the Social Sciences? *American Sociological Review*, 69(2), 190-212.
Hammarfelt, B. (2012). Harvesting footnotes in a rural field: citation patterns in Swedish literary studies. *Journal of Documentation*, 68(4), 536-558.
Hammarfelt, B. (2016). Beyond Coverage: Toward a Bibliometrics for the Humanities. In M. Ochsner, S. E. Hug & H.-D. Daniel (pp. 115-131). *Research Assessment in the Humanities: Towards Criteria and Procedures*. SpringerOpen.
Hargens, L. L. (2000). Using the literature: Reference networks, reference contexts, and the social structure of scholarship. *American Sociological Review*, 65(6), 846-865.
Harwood, N. (2009). An interview-based study of the functions of citations in academic writing across two disciplines. *Journal of Pragmatics*, 41(3), 497-518.
Heinzkill, R. (1980). Characteristics of References in Selected Scholarly English Literary Journals. *Library Quarterly*, 50(3), 352-365.
Hellqvist, B. (2010). Referencing in the humanities and its implications for citation analysis. *Journal of the American Society for Information Science and Technology*, 61(2), 310-318.
Hicks, D. (1999). The difficulty of achieving full coverage of international social science literature and the bibliometric consequences. *Scientometrics*, 44(2), 193-215.
Hicks, D. (2004). The Four Literatures of Social Science. In H. F. Moed, W. Glänzel & U. Schmoch (Eds.). *Handbook of Quantitative Science and Technology Research* (pp 473-495). Dordrecht: Kluwer.
Hyland, K. (1999). Academic attribution: citation and the construction of disciplinary knowledge. *Applied Linguistics*, 20(3), 341-367.
Journal of the International Association of Research Institute in the History of Arts (2019). *Guidelines for authors*. Retrieved January 25, 2019 from https://www.riha-journal.org/guidelines/for-authors.
Kristensen, P. M. (2018). International Relations at the End: A Sociological Autopsy. *International Studies Quarterly*, 62(2), 245-259.

Linmans, A. J. M. (2010). Why with bibliometrics the Humanities does not need to be the weakest link. *Scientometrics*, 83(2), 337-354.
Moed, H. F., Luwel, M. & Nederhof, A. J. (2002). Towards Research Performance in the Humanities. *Library Trends*, 50(3), 498-520.
Nederhof, A. J. (2006). Bibliometric monitoring of research performance in the Social Sciences and the Humanities: A review. *Scientometrics*, 66(1), 81-100.
Oltersdorf, J. (2013). Publikationen: Funktion und Repräsentation – Präsenz von Kommunikationskanälen der deutschen Kunstgeschichte in bibliographischen Nachweisinstrumenten. PhD, Humboldt-University of Berlin. Retrieved January 25, 2019 from https://edoc.hu-berlin.de/bitstream/handle/18452/17443/oltersdorf.pdf?sequence=1&isAllowed=y.
Panofsky, E. (1955). *Meaning in the Visual Arts.* Garden City: Doubleday.
Shadish, W. R., Tolliver, D., Gray, M. & Gupta, S. K. S. (1995). Author Judgements about Works They Cite: Three Studies from Psychology Journals. *Social Studies of Science*, 25(3), 477-498.
Sivertsen, G. (2016). Patterns of internationalization and criteria for research assessment in the social sciences and humanities. *Scientometrics*, 107(2): 357-368.
Stern, M. (1983). Characteristics of the Literature of Literary Scholarship. *College & Research Libraries*, 44, 199-209.
Sternberg, R. & Litzenberger, T. (2005). The publication and citation output of German faculties and disciplines based upon SSCI data. *Scientometrics*, 65, 29-53.
Van Leeuwen, T. N. (2006). The application of bibliometric analyses in the evaluation of social science research. Who benefits from it, and why it is still feasible. *Scientometrics*, 66(1), 133-154.
Van Leeuwen, T. N. (2013). Bibliometric research evaluations, Web of Science and the Social Sciences and Humanities: a problematic relationship? *Bibliometrie – Praxis und Forschung*, 2, 8/1 - 8/18.
Van Raan, A. F. J. (1996). Advanced bibliometric methods as quantitative core of peer review based evaluation and foresight exercises. *Scientometrics*, 36(3), 397-420.
White, H. D., Boell, S. K., Yu, H., Davis, M., Wilson, C. S. & Cole, F. T. H. (2009). Libcitations: A measure for comparative assessment of book publications in the humanities and social sciences. *Journal of the American Society for Information Science and Technology*, 60(6), 1083-1096.
Whitley, R. (2000 [1984]). *The Intellectual and Social Organization of the Sciences.* Oxford, Clarendon Press.
Zuccala, A. & Cornacchia, R. (2016). Data matching, integration, and interoperability for a metric assessment of monographs. *Scientometrics*, 108(1), 465-484.

---

[i] Another, less elaborate attempt is Hammarfelt's (2012) account of Swedish literary studies as a "rural" field as described by Becher and Trowler (Becher and Trowler 2001).

[ii] CIHA was founded in 1873 and is the oldest international organisation of art history in the world. One of its main aims is to foster cooperation between art historians of all countries by encouraging international meetings and the dissemination of information about art history research.

[iii] "Whoever comments on art and art history today sees any thesis they want to communicate to the reader devalued *ex ante* by uncountable other theses." (Belting 1995: 17, our translation)

[iv] This entails an interesting asymmetry because in order to fully reconstruct the context of a work of art, art historians must access a wide range of sources in other language than their mother tongue. The use of these sources and the use of their foreign colleagues' publications cannot occur at the same level of understanding and precision as their own production. This asymmetry needs further investigation.

Does collaborative research published in top journals remain uncited?

A. I. M. Jakaria Rahman

jakaria.rahman@chalmers.se
Department of Communication and Learning in Science, Chalmers University of Technology
SE-41296 Gothenburg, (Sweden)

Abstract

This paper investigates whether collaborative research published in top journals remains uncited, and to what extent access type (open and closed) affects on citation of collaborative research published in top journals. It looks at publications including articles, conference papers, reviews, short surveys, editorials, letters, notes published between 2009-2016 with an affiliation from Chalmers University of Technology and indexed in Scopus. To give enough time to gather citation, two-year time frame is considered for the publication of the year 2016. The data is classified based on access types: closed and open access, and sub-classified as cited closed access, cited open access, non-cited closed access, and non-cited open access in SciVal. The top 25 percentile indicating the number of journals that are in the top 25% of the most cited journals indexed by Scopus is considered. The result shows that a small portion of collaborative research published in top journals remain uncited irrespective of types of collaboration. In case of international collaborative research, publications in closed access are more cited than in open access. Institutional collaborative research publications are more cited than national collaborative ones. Collaborative research is more cited than single authors' publications and single authored conference papers published in the top percentile do not remain uncited.

Introduction

It is obvious that researchers would be happy to receive a citation instead of simply publishing. Some academic publications receive citation immediately after publication, while some are cited within 2-5 years, and some are never cited, and the percentage of non-cited publications vary by discipline (Burrell, 2002; Hu, Wu, & Sun, 2018; Liang, Zhong, & Rousseau, 2015; Van Noorden, 2017) and document types i.e. articles, conference papers, review articles, letters, and notes (Cano & Lind, 1991) book, book chapters, etc. (Bott & Hargens, 1991). Non-cited publication neither mean that there is no chance of being cited in the future, nor that the publication has never been read. Some publications may take many years to be recognized and to receive citations (Braun, Glänzel, & Schubert, 2010; Ke, Ferrara, Radicchi, & Flammini, 2015; van Raan, 2004, 2017; Zong, Liu, & Fang, 2018). If a publication is not cited in a citation database, for example, Scopus or Web of Science, it does not mean that it is uncited. It could be cited and available in other places, for example, google scholar. Simultaneously, it could be possible that a publication was uncited while the data was retrieved (Liang et al., 2015). Concurrently, funding agencies, promotion and recruitment committees use citation data to evaluate a researcher in addition to peer review evaluation (Meho, 2007) as well as the university ranking organizations (Waltman et al., 2012).

High quality research receive more citation than low quality research (Meho, 2007; van Raan, Visser, Van Leeuwen, & van Wijk, 2003). There are several factors like number of authors, title words, keywords, number of references, journal age, price, etc. that influence a publication being cited (Stern, 1990). Poor knowledge, lack of original contents, late discovery, bibliographic plagiarism, academic misconduct, etc. are listed by Garfield (1991) as reasons for a publication to remain uncited. Rousseau (1992) observes that multi-authored publications receive more citation than single authored ones. Journal impact factor, journal's age, average number of references per paper, number of issues of a journal also have influence on citation (Hu et al., 2018). On the other hand, it is quite common that many renowned scientists including Nobel laureates have uncited publications (Egghe, Guns, & Rousseau, 2011). Different number

of citations for different publications of a single author indicates the varying quality of the publications (Burrell, 2012).

Publication with international collaboration (international co-authorship) receives higher citation than national collaboration (single country inter-institutional co-authorship), and two times higher than institutional collaboration (single-country single-institution co-authorship) (Narin, Stevens, & Whitlow, 1991; Smith, Weinberger, Bruna, & Allesina, 2014). There is influence of the publication and collaboration habits in the different field (Coccia & Wang, 2016) and research productivity is influenced by the collaboration habits (Abramo, D'Angelo, & Di Costa, 2019). As far our knowledge, citation rate of collaborative research published in the top journals and effect of access types of those publications in receiving citation are still unexplored. Therefore, collaborative research published in top journals remain uncited or not is a legitimate object of research. Hence, we explored two research questions:

i) To what extent collaborative research published in top journals remained uncited?

ii) To what extent access type (closed and open) affected citation of collaborative research published in top journals?

Data and method

As a test case, we considered Chalmers University of Technology's (here after Chalmers) publications that were indexed in Scopus database. Chalmers conducts research in technology, science, architecture and maritime engineering. On an average, around 2200 peer reviewed publications contributed by Chalmers scholars per year were indexed by Scopus. The publications considered here, more specifically, were articles, conference papers, reviews, short surveys, editorials, letters, notes that were published in an 8-year window (2009-2016). The publication year 2017 and 2018 were not included to give enough time to gather citation for the publications of the last year (2016) considered. We downloaded the data with Scopus's own digital identifiers known as EID. As SciVal produces bibliometric analysis based on Scopus, we limited our dataset to Scopus only.

We found 17917 publications with Chalmers affiliation during the 8-years window. The data set was further classified based on number of times a publication has received citation. If a publication received at least one citation (including self-citation), we classified it as cited, otherwise non-cited. In addition, we also classified based on access types, i.e. open access and closed access. Altogether, we created four publication data set namely, cited closed access, cited open access, non-cited closed access, and non-cited open access using the corresponding EIDs (Table 1).

We explored the collaboration types as it indicates the extent to which a publication is of international, national or institutional co-authorship, and single authorship. When at least two authors from two different organizations co-authored in an article, we considered it as a collaboration. In the case of single authorship, we considered no collaboration even the author might have two affiliations. We used the default set up of SciVal to handle our data set for calculating collaboration.

For further analysis, we put articles, reviews, editorials and short surveys in one cluster and the conference papers in another cluster (see Table 2). We created these two-clusters due to default set up of SciVal for benchmarking. The first cluster covered the major documents types (excluding letter and note) and the second cluster covered all the conference papers.

Table 1 Publication profile of Chalmers University of Technology from 2009 – 2016 in Scopus.

Document type	*Closed access*	*Cited* *Open Access*	*Total*	*Closed access*	*Non-cited* *Open Access*	*Total*	*Grand Total*
	a	*b*	*c = a + b*	*d*	*e*	*f = d + e*	*g = c + f*
Article	8054 (74%)	2387 (21%)	10441 (95%)	423 (4%)	81 (1%)	504 (5%)	10945
Conference Paper	4214 (67%)	327 (5%)	4541 (72%)	1659 (26%)	99 (2%)	1758 (28%)	6299
Editorial	28 (19%)	21 (14%)	49 (34%)	63 (43%)	34 (23%)	97 (66%)	146
Letter	19 (56%)	9 (26%)	28 (82%)	4 (12%)	2 (6%)	6 (18%)	34
Note	22 (42%)	16 (30%)	38 (72%)	11 (21%)	4 (8%)	15 (28%)	53
Review	286 (69%)	107 (26%)	393 (95%)	19 (5%)	–	19 (5%)	412
Short Survey	14 (50%)	12 (43%)	26 (93%)	2 (7%)	–	2 (7%)	28
Grand total	12637 (71%)	2879 (16%)	15516 (87%)	2181 (12%)	220 (1%)	2401 (13%)	17917

This default cluster in SciVal is one of the limitations to not include letters and notes for this investigation while calculating collaboration. To cover this limitation, we considered all the document types (all publication in SciVal) too as a part of analysis to see how the entire set of data responded in the analysis.

The publications in top journal percentiles indicates the number of publications that have been published in the world's top journals (the most-cited journals indexed by Scopus database). Here, we considered the top 25 percentile that indicates the number of journals that are in the top 25% of the most cited journals indexed by Scopus. In SciVal, we can also find top 1%, 5% and 10% journal percentile. The top 25 percentile covers all the share of the top percentiles. Therefore, the focus was given on only the journals that are in the top 25 percentile where Chalmers has publications.

The most-cited journals are defined by the journal metrics that have a CiteScore percentile (using citation data from the Scopus database to rank journals), SNIP (Source-Normalized Impact per Paper) or SJR (SCImago Journal Rank) (Elsevier, 2018). Fields with low citation numbers are penalised in CiteScore and SJR ranks publications by weighted citations per document and weighting depends on subject field and prestige of the citing publications. (Elsevier, 2018). While "SNIP corrects for differences in citation practices between scientific fields, thereby allowing for more accurate between-field comparisons of citation impact" – stated by the Centre for Science and Technology Studies (CWTS, 2019). Therefore, we choose SNIP value for retrieving data as we focused on the top 25 percentile journals. We analysed the data based on international collaboration, national collaboration, institutional collaboration and single authorship. A publication could fall in international collaboration, national collaboration, institutional collaboration through its affiliation.

Table 2 Comparison between publications in the top 25 percentile journals and collaboration.

Category	*Cluster 1: Articles, reviews, editorials, short surveys*					*Cluster2: Conference papers*				
	Top 25% of Scopus Source	*International Collaboration*	*National Collaboration*	*Institutional Collaboration*	*Single author*	*Top 25% of Scopus Source*	*International Collaboration*	*National Collaboration*	*Institutional Collaboration*	*Single author*
All publications	8199 (95%)	4704 (57%)	1361 (17%)	1760 (22%)	374 (5%)	419 (5%)	168 (40%)	113 (27%)	128 (31%)	10 (2%)
Cited	7998 (98%)	4591 (56%)	1336 (17%)	1731 (22%)	340 (4%)	382 (91%)	153 (37%)	104 (25%)	115 (27%)	10 (2.4%)
Cited closed access	6108 (74%)	3257 (40%)	1071 (13%)	1509 (18%)	271 (3%)	303 (72%)	139 (33%)	68 (16%)	87 (21%)	9 (2%)
Cited open access	1890 (23%)	1334 (16%)	265 (3%)	222 (3%)	69 (0.8%)	79 (19%)	14 (3%)	36 (9%)	28 (7%)	1 (0.2%)
Non-cited	201 (2%)	113 (1%)	25 (0.3%)	29 (0.3%)	34 (0.4%)	37 (9%)	15 (4%)	9 (2%)	13 (3%)	0
Non-cited closed access	156 (2%)	87 (1%)	19 (0.2%)	25 (0.3%)	25 (0.3%)	28 (7%)	15 (4%)	3 (0.7%)	10 (2%)	0
Non-cited open access	45 (0.5%)	26 (0.3%)	6 (0.07%)	4 (0.05%)	30 (0.1%)	9 (2%)	0	6 (1%)	3 (0.7%)	0

Further, we classified all the publications that fall in the top 25 percentile of the Scopus source in two major categories: cited and non-cited, and both the categories were sub-categorised as closed access and open access (Table 2).

Analysis and Results

We found 15516 (87%) publications with at least one citation (including self-citation) whereas 2401 (13%) publications remained uncited from a range of 2 to 9 years (see Table 1). Articles (95%), reviews (95%) and short surveys (93%) were mostly cited. Closed access (71%) publications were more cited than the open access (16%) publications, while in the case of non-cited publication, open access publications (1%) were less than the closed access (12%) publications. A small percentage (5%) of the articles did not received citation while it is 28% for the conference papers, editorials (66%), letters (18%), notes (28%), reviews (5%), short surveys (7%). Other than editorials, the larger share of cited and non-cited publications were published as closed access.

A larger share of conference papers (28%) remained uncited where the majority was published in closed access (26%). Editorials (66%) were the largest share among the non-cited document types that would not usually get cited (Van Noorden, 2017). According to Scopus (2017), review articles were defined as a 'significant review of original research' and considered short surveys similar to reviews (but usually are shorter). We found that none of these two document types remained uncited while published as open access.

We found 1112 (6%) publications were single authored and 8618 (48%) publications were published in the top 25% journals out of 17917 publications. Table 2 indicates that in cluster one (Articles, reviews, editorials, short surveys), 8199 (95%) publications published in the

world's top 25% journals, while only 419 (5%) for the cluster two (Conference papers). In cluster one, 98% cited publication and 2% non-cited publications fall in the top 25% journals. Further, international collaboration (56%), national collaboration (17%) and institutional collaboration (22%) were higher in cited publications than non-cited publications (1%, 0.3% and 0.3% respectively). In addition, single author cited publications (4%) was higher than single author non-cited publications (0.4%).

Cited open access had less international collaboration (16%) than cited closed access (40%). Similarly, non-cited closed access publications (1%) were higher than non-cited open access (0.3%). At the same time, national collaboration in cited closed access publications (13%) was higher than cited open access publications (3%). Institutional collaborative research (22%) is more cited than national collaborative research (17%).

Table 2 also indicates that in cluster two (conference papers), the percentage of publication in the top 25% were very low (5%) in both cited and non-cited categories. Remember that the Table 1 indicates that conference paper was the second large set in our retrieved data. The international collaboration (33%) was higher than the national (16%) and institutional (27%) collaboration in cited closed access but lower in cited open access. Both in cited and non-cited cases, institutional collaboration (21% and 3%) was higher than the national collaboration (16% and 2%) for conference papers. All the single authored conference papers received citation.

Discussion and Conclusion

In the dataset, 13% publications remained uncited from a range of 2 to 9 years. That is, only a small portion of the articles, reviews and short surveys remained uncited. As a Science and technology university, most of the departments of Chalmers participate in their respective flagship conferences but nearly one-third of the conference papers remain uncited. In many cases, conference papers are foundations to the creation of journal articles (Drott, 1995). While recent conference papers are primary source of new research as needed by technology related researchers, these are less citable when older (Young, 2014). Based on research field or discipline, the same document type takes different meaning, and even the journals of the same discipline have different structure and distinguish document types (Sigogneau, 2000). In our data set, most cited and non-cited publications were published as closed access. Further, the closed accessed publications were more cited than open access which supports the finding of Craig, Plume, McVeigh, Pringle, & Amin (2007). In addition, most of the uncited conference papers were published as closed access. However, reviews and short surveys were more cited while published as closed access.

We conclude that a small portion of collaborative research published in top journals remain uncited irrespective of types of collaboration. Collaborative research is more cited than single author publications. In our data set, closed access publications are cited more than open access publications. Publication in open access top journals are less cited than closed access top journals while the research is collaborated internationally.

This paper includes a data set which is less than 20 thousand publications due to the limitation of Scopus and SciVal data handling process. An investigation with wider window like 10- or 20-years data from both 'science and technology' and 'social science and humanities' and not just within an institution's publications also different open access publishing models (green, gold or hybrid), and additional data sources like Web of Science citation index and Google scholar will facilitate to retrieve information that might allow the result to generalize. In addition, breaking down the various indicators by disciplinary and sub-disciplinary categories,

and the country of publication of the journals may provide insightful information about why some collaborative research published in top journals remain uncited.

Acknowledgment

The author thanks three anonymous reviewers for their constructive comments, and Hasan Mahmud and Momena Khatun for suggestions.

References

Abramo, G., D'Angelo, C. A., & Di Costa, F. (2019). The collaboration behavior of top scientists. *Scientometrics*, *118*(1), 215–232.

Bott, D. M., & Hargens, L. L. (1991). Are sociologists' publications uncited? Citation rates of journal articles, chapters, and books. *The American Sociologist*, *22*(2), 147–158.

Braun, T., Glänzel, W., & Schubert, A. (2010). On Sleeping Beauties, Princes and other tales of citation distributions …. *Research Evaluation*, *19*(3), 195–202.

Burrell, Q. L. (2002). Will this paper ever be cited? *Journal of the American Society for Information Science and Technology*, *53*(3), 232–235.

Burrell, Q. L. (2012). Alternative thoughts on uncitedness. *Journal of the American Society for Information Science and Technology*, *63*(7), 1466–1470.

Cano, V., & Lind, N. C. (1991). Citation life cycles of ten citation classics. *Scientometrics*, *22*(2), 297–312.

Coccia, M., & Wang, L. (2016). Evolution and convergence of the patterns of international scientific collaboration. *Proceedings of the National Academy of Sciences*, *113*(8), 2057–2061.

Craig, I. D., Plume, A. M., McVeigh, M. E., Pringle, J., & Amin, M. (2007). Do open access articles have greater citation impact?: A critical review of the literature. *Journal of Informetrics*, *1*(3), 239–248.

CWTS. (2019). Welcome to Centre for Science and Technology Studies journal indicators. Retrieved January 25, 2019, from http://www.journalindicators.com

Drott, M. C. (1995). Reexamining the role of conference papers in scholarly communication. *Journal of the American Society for Information Science*, *46*(4), 299–305.

Egghe, L., Guns, R., & Rousseau, R. (2011). Thoughts on uncitedness: Nobel laureates and Fields medalists as case studies. *Journal of the American Society for Information Science and Technology*, *62*(8), 1637–1644.

Elsevier. (2018). *Research metrics guidebook*. Retrieved from https://www.elsevier.com/research-intelligence/resource-library/research-metrics-guidebook

Garfield, E. (1991). To be an uncited scientist is no cause for shame. *Scientist*, *5*(6).

Hu, Z., Wu, Y., & Sun, J. (2018). A quantitative analysis of determinants of non-citation using a panel data model. *Scientometrics*, *116*(2), 843–861.

Ke, Q., Ferrara, E., Radicchi, F., & Flammini, A. (2015). Defining and identifying Sleeping Beauties in science. *Proceedings of the National Academy of Sciences of the United States of America*, *112*(24), 7426–7431.

Liang, L., Zhong, Z., & Rousseau, R. (2015). Uncited papers, uncited authors and uncited topics: A case study in library and information science. *Journal of Informetrics*, *9*(1), 50–58.

Meho, L. I. (2007). The rise and rise of citation analysis. *Physics World*, *20*(1), 32–36.

Narin, F., Stevens, K., & Whitlow, E. S. (1991). Scientific co-operation in Europe and the citation of multinationally authored papers. *Scientometrics*, *21*(3), 313–323.

Rousseau, R. (1992). Letter to the editor. *Journal of Documentation*, *48*(1), 79–80.

Scopus. (2017). *Scopus: content coverage guide*. Retrieved from http://www.elsevier.com/solutions/scopus/content

Sigogneau, A. (2000). An analysis of document types published in journals related to Physics: Proceeding papers recorded in the Science citation index database. *Scientometrics*, *47*(3), 589–604.

Smith, M. J., Weinberger, C., Bruna, E. M., & Allesina, S. (2014). The Scientific impact of nations: Journal placement and citation performance. *PLoS ONE*, *9*(10), e109195.

Stern, R. E. (1990). Uncitedness in the biomedical literature. *Journal of the American Society for*

Information Science, *41*(3), 193–196.
Van Noorden, R. (2017). The science that's never been cited. *Nature*, *552*(7684), 162–164.
van Raan, A. F. J. (2004). Sleeping Beauties in science. *Scientometrics*, *59*(3), 467–472.
van Raan, A. F. J. (2017). Sleeping beauties cited in patents: Is there also a dormitory of inventions? *Scientometrics*, *110*(3), 1123–1156.
van Raan, A. F. J., Visser, M. S., Van Leeuwen, T. N., & van Wijk, E. (2003). Bibliometric analysis of psychotherapy research: Performance assessment and position in the journal landscape. *Journal Landscape, Psychotherapy Research*, *13*(4), 511–528.
Waltman, L., Calero-Medina, C., Kosten, J., Noyons, E. C. M., Tijssen, R. J. W., van Eck, N. J., … Wouters, P. (2012). The Leiden ranking 2011/2012: Data collection, indicators, and interpretation. *Journal of the American Society for Information Science and Technology*, *63*(12), 2419–2432.
Young, B. (2014). What do Engineering researchers cite? A citation analysis study of sixteen Engineering journals. *Issues in Science and Technology Librarianship*, *75*.
Zong, Z., Liu, X., & Fang, H. (2018). Sleeping beauties with no prince based on the co-citation criterion. *Scientometrics*, *117*(3), 1841–1852.

Using a local database to uncover non-source items: the case of Science Education in Brazil using the Sucupira Platform

Eloisa Viggiani[1] and Luciana Calabró[2]

[1] *maileloisa@gmail.com*
Universidade Federal do Rio Grande do Sul, Rua Ramiro Barcelos 2600, 90035-003, Porto Alegre (Brazil)

[2] *luciana.calabro@ufrgs.br*
Universidade Federal do Rio Grande do Sul, Rua Ramiro Barcelos 2600, 90035-003, Porto Alegre (Brazil)

Abstract

Research in the Social Sciences and Humanities is published in a variety of formats and types of documents. Local journals are relevant for these fields, particularly in non-English-speaking countries, but coverage of large bibliographic databases is still limited. This represents a challenge for Performance-based Research Funding Systems and also for national evaluation exercises, such as the evaluation of the graduate degree programs performed by CAPES in Brazil, registered in the Sucupira Platform. The purpose of this article is to explore the use of Sucupira to analyse the research output not covered by WoS and Scopus. We extracted the publications reported by 140 degree programs from the Teaching area -also known as Science Education, and identified 5,255 non-source articles published in 162 journals not indexed in WoS or Scopus. Our study shows that Sucupira is an essential source of data to obtain a comprehensive account of Brazil´s research activity and output in Humanities, Social Sciences and multidisciplinary areas, and should be used in addition to WoS and/or Scopus in large-scale, national bibliometric studies.

Introduction

There is a general understanding that researchers in the Humanities and Social Sciences tend to publish the results of their research in a variety of formats, such as journal articles, books and book chapters, conference proceedings, reports and other types of documents. Local journals tend to be relevant for these fields, particularly in the non-English-speaking countries (Hicks 1999 & 2010, Moed 2005). Even though both Web of Sciences (WoS) and Scopus have made efforts to increase their coverage of non-English sources in the Humanities and Social Sciences (Collazo-Reyes 2013, Dagiene 2015), many local journals are not indexed by these bibliographic databases. Articles published in these journals have been referred to as "non-source" publications by the literature (Butler 2006, Boyak 2014, Chi 2014) and can represent a considerable proportion of the research output of non-English-speaking countries.

This results in a challenge for Performance-based Research Funding Systems - PRFS (Hicks, 2012) and for this reason, many countries have adopted peer-review systems to develop evaluation criteria for publications not indexed in the bibliographic databases. This is the case of Brazil, where a national evaluation of the graduation degree programs takes place regularly, conducted by CAPES, the Federal Agency for Support and Evaluation of Graduate Education, under the Ministry of Education (CAPES, 2017d).

The main purpose of the evaluation exercise is to ensure the quality of degree programs (professional masters, academic masters and PhDs). The results by program may be aggregated to the university level in order to plan government funding for universities, award grants and provide research resources, such as access to the CAPES Portal, a national electronic library consortium (CAPES, 2019), but not all government distribution of research funding in the country depends on the results of the evaluation. These characteristics distinguish the CAPES evaluation from Hicks´ definition of PRFS, nevertheless, there are many commonalities: it is a national system, where an ex-post research output evaluation is regularly and consistently performed across all research fields and, to assist this evaluation, the Ministry of Education implemented a criterion system for classifying journals, the Qualis system (CAPES, 2017c).

Additionally, the national evaluation becomes an important instrument to understand Brazil´s research performance because several studies show that the practice of research in the country is strongly concentrated within the graduate programs (De Meis 2007, Coutinho 2012), According to Leta (2006), universities from the public sector account for more than 80% of the country´s total publications in the WoS database.
The criteria and results of the national evaluation conducted by CAPES are made publicly available through an online platform, namely the Sucupira Platform, a recent improvement in the process of collecting data and performing analysis (CAPES, 2014). Indicator Spreadsheets with data regarding the teaching and research human resources, activities and outputs reported by the programs can be downloaded from Sucupira.
The objective of this study is to explore the potential use of Sucupira Platform to analyze the research output not covered by the major bibliographic databases. We selected one of the 49 CAPES evaluation areas, the Teaching area, also known as Science Education. By definition, the Teaching area aims the integration of knowledge from the Education area and several other disciplinary fields (CAPES, 2017a). This area was selected because of the low coverage of Brazilian journals in Education in major databases (Madrid-Martín, 2017) and the fact that research trends in Science Education in Latin America are uncommon in the literature (Medina-Jerez, 2018), with one of the few exceptions being the study by Coutinho et al. (2012) about the Brazilian scientific production in Science Education.

Background

National Evaluation of the Graduate Degree Programs in Brazil

The national evaluation of the graduate degree programs is currently performed every 4 years and the latest evaluation exercise was performed in 2017, covering the period between 2013 and 2016. The results were made available by CAPES in September 2017 (CAPES, 2017b).
The evaluation framework uses a subject area classification system with 49 different Evaluation Areas, grouped into 9 major areas. Each program is assigned to an Evaluation Area, as can be observed in Table 1.

Table 1. Examples of program assignment to CAPES´ Subject Area Classification.

Master and/or PhD Program	*2nd Tier Evaluation Area (*)*	*1st Tier Major Areas (*)*
Applied Physics	Astronomy/Physics	Exact & Earth Sciences
Biochemistry	Biological Sciences II	Biological Sciences
Education and Diversity	Education	Humanities
Physics Teaching	Teaching	Multidisciplinary

(*) CAPES Subject Area Classification: the 1st Tier contains 9 Major Areas and the 2nd Tier contains 49 Evaluation Areas to which every graduation program is assigned (CAPES, 2017)**.**

Each evaluation area has a committee of experts and ad-hoc consultants, who have the discretion to determine evaluation criteria, such as how to classify the peer-reviewed journals in which the programs belonging to their area have reported publications. This is known as the Qualis journal classification, which is unique for each evaluation area (CAPES, 2017c). For example, the areas Astronomy/Physics and Biological Sciences adopt the Impact Factor (Clarivate Analytics, 2019) as the main quantitative indicator for journal classification, were as the Education and the Teaching area prioritize a combination of the journal´s disciplinary focus, relevance for the area and indexation in a broader set of databases (CAPES, 2017a). This means

that a journal may receive different level classifications depending on the evaluation area, as can be observed in Table 2.

Table 2. Examples of Qualis journal classification levels attributed to journal titles by different Evaluation Areas.

Journal Title and ISSN	*PLOS One 1932-6203*	*Science & Education 0926-7220*	*Physical Review Letters 0031-9007*
Evaluation Area	*Qualis classification level attributed to journal*		
Astronomy/Physics	A1	B5	A1
Biological Sciences II	B1	-	A1
Education	A2	A1	-
Teaching	B2	A1	B2

Scoring: A1 = 100, A2 = 85, B1 = 70, B2 = 55, B3 = 40, B4 = 25, B5 =10.

Titles are classified as A1, A2, B1, B2, B3, B4, B5 or C, where A1 and A2 are the top levels and C means "no value" (CAPES, 2017c). If a journal has no publications reported by an Evaluation area during the evaluation period, it will not figure in the area´s Qualis Journal Classification list.

These publications, together with publications in books and selected conferences, technical and artistic productions constitute the Intellectual Production of the program, which represents between 20% - 40% of the total scoring for professional Master programs and 35% - 40% for academic Master and PhD programs.

The Sucupira Platform

Launched in 2014, the Sucupira platform succeeded the Coleta Platform and became the online tool to collect data and conduct analysis for the national evaluation of the graduation degree programs in Brazil (CAPES, 2014). Program Coordinators are responsible for inputting their teaching and research activities into Sucupira. The starting point is registering all program faculty members via the individual taxpayer registry identifier CPF, so the system will validate name and other personal data. This allows data to be entered manually or imported from each faculty´s curriculum in the Lattes Platform, a national database of résumés maintained by the National Council for Scientific and Technological Development – CNPq (CNPq, 2019). In 2016, the Lattes database contained résumés of 132,831 PhDs and 82,818 Masters dedicated to teaching and research, as well as 86,831 PhDs and 281,822 Masters dedicated to administrative, technical and other activities (CNPq, 2016). Lattes also uses the CPF as the main personal identifier and once the Lattes curriculum is located, the Program Coordinator may select one or more publications authored by a faculty member to be imported into Sucupira by a native application. Besides the author(s), other mandatory fields for inputting articles as research outputs in Sucupira are journal ISSN, volume, language (of the article) and media (print or electronic). The article DOI is a text field, non-mandatory.

The results of the evaluation are made publicly available through the Sucupira Platform. For each evaluation area an Indicator Spreadsheet can be downloaded, containing data about teaching and research human resources, activities and outputs reported by the programs.

The Teaching Area

The main objective of the programs belonging to the Teaching area is the training of masters and doctors for the teaching of Sciences and Mathematics, as we can observe in the specific

objective of the Professional Master in Physics Teaching at the Federal University of Rio Grande do Sul - UFRGS:

> "Improvement of the professional qualification of Physics teachers… in terms of Physics contents, theoretical, methodological, epistemological aspects of Physics teaching, and the use of new technologies in Physics teaching." *(UFRGS, 2018)*

Research in the Teaching area deals with educational processes in an interdisciplinary way and the focus is on the integration of disciplinary content and pedagogical knowledge, and is many times referred to as Science Education. Regarding research output, the Teaching area values a combination of high international visibility and disciplinary focus on Education, Teaching, Cognition and/or Learning (ETCL). This is reflected in the area´s Qualis journal classification, where the highest classification levels are assigned to journals specialized in ETCL, followed by multidisciplinary, discipline specific related to ETCL and discipline specific not related to ETCL, as illustrated in Table 3.

Table 3. Criteria for Qualis Journal Classification for the Teaching Area.

	Journal disciplinary focus			
Covered by	*Specialized In ETCL*	*Multidisciplinary*	*Related Discipline*	*Unrelated Discipline*
WoS, Scopus or SciELO	A1			
WoS or Scopus		A2	B1	
SciELO		B1	B2	
ERIC, DOAJ or Latindex	B1	B2	B3	
Google Scholar w/ h5 index	A2	B3	B4	
Google or other	B2	B5		
WoS or Scopus, IF ≥1.5				B2
WoS or Scopus, 1.0≤IF<1,5				B3
WoS or Scopus, 0,5≤IF<1.0				B4
Any database		B4	B5	B5
Available in schools	B3			

Scoring: A1 = 100, A2 = 85, B1 = 70, B2 = 55, B3 = 40, B4 = 25, B5 =10
ETCL = Education, Teaching, Cognition and/or Learning

Journals classified as A1 should be specialized in ETCL and be indexed by either WoS, Scopus or SciELO, so in the Teaching area, the most valued journals are expected to be field-specific and have high international visibility. Therefore, we expect to find the lowest number of non-source journals in A1, with the only exception being the ones indexed by SciELO but not by WoS or Scopus. Journals classified as A2 can be either multidisciplinary indexed by WoS or Scopus, or ETCL-specific but not indexed in WoS, Scopus or SciELO, as for example the journal *Revista Brasileira de Ensino de Ciência e Tecnologia* - RBECT, whose mission to:

> "... disseminate practical and theoretical research on the teaching-learning process, resulting from a reflexive, critical and innovative action for the teacher´s work, support the knowledge production and new pedagogical strategies in Science, Biology, Chemistry, Physics, Mathematics, Engineering and Technology teaching." *(RBECT, 2019)*

We expect to find the highest number of non-source journals in the classifications A2 to B4, where journals may (or not) be covered by SciELO, ERIC, DOAJ, Latindex, Google Scholar

or any other database. In the B3 classification we can find non peer-reviewed journals that are specialized in ETCL and available in schools, such as *Ciência Hoje das Crianças*, first Brazilian journal dedicated to children and currently distributed to 60,000 school libraries in the country (CHC, 2019). Journals classified as B5, the lowest rank, are not expected to be covered by WoS or Scopus, but still this could be the case when the journal is specialized in a disciplinary field unrelated to ETCL.

Methodology

We considered information from: articles published by faculty members in scholarly journals of the 140 Masters and PhDs degree programs from the Teaching area, which were submitted to the quadrennial evaluation in 2017 (CAPES, 2017b). This was obtained through the Sucupira Platform, downloading the Indicator Spreadsheet for the Teaching area at http://avaliacaoquadrienal.capes.gov.br/home/planilhas-de-indicadores . The sheet contains the most recent data, updated on September 28, 2017 (CAPES, 2017d), The information is arranged in 11 tabs, of which 4 were consulted:

- Programs: contains a list of all programs with faculty size, summary of activities, research projects and thesis & dissertations
- Output list: contains a list of articles published by each program per year, with article name, journal title, ISSN and Qualis ranking
- Output per program: contains the number of articles published by each program per year, with number of author (students and faculty members) and Qualis ranking.

From the "Output list" tab we obtained a list of all publications in scholarly journals reported by the programs from the Teaching area, published between 2013 and 2016. A semi-manual process was employed to remove duplicate records of articles, which were reported more than once due to co-authorship between programs or due to name variants and typos in journal and/or publication titles. The result was a set of unique publications, which was aggregated by journal´s ISSN number. We then created a list of journals with the number of publications reported by the Teaching area programs. This list was sorted and to avoid a "long-tail" effect all journals that had less than 10 publications between 2013-2016 were considered not relevant for the Teaching area and were discarded from the analysis. Journals classified as "C" -which means "no value"- were also discarded.

The remaining list of journals was compared via the journals´ ISSN number to the lists of titles covered by all the databases and search engines adopted in the Qualis classification criteria for the Teaching area, namely: WoS (Clarivate Analytics, 2018), Scopus (Elsevier, 2018), SciELO (SciELO , 2018), ERIC (ERIC, 2019), DOAJ (DOAJ, 2019), Latindex (Latindex, 2019), and Google Scholar (Google Scholar, 2019). We obtained the journal coverage from their websites. Our search in WoS was limited to journals covered by the Core Collection (Science Citation Index, Arts and Humanities Citation Index, Social Sciences Citation Index and Emerging Sources Citation Index) and in Google Scholar limited to journals with an h5 index. The number of titles covered by each database was calculated per classification level.

Of special interest are the two main bibliographic databases used for evaluation purposes, WoS and Scopus. For the purpose of this article, we adapted the definitions proposed by Butler (2006) and Chi (2013) to the following:

- Non-source titles: journals not covered by Scopus or WoS Core Collection.
- Non-source publications: articles in journals not covered by Scopus or WoS Core Collection.

Our definition of non-source journals and publications has not considered other databases and search engines for a variety of reasons: though the SciELO database registers citations to its

articles, it is not likely to be used for a large scale bibliometric study because its coverage is limited to Open-access journals in Latin America, the Caribbean, Spain and Portugal; DOAJ does not have geographical limitations but is also restricted to Open-access journals; ERIC is highly selective and only covers materials in English language, and Latindex does not contain citation metrics; Google Scholar has a broad coverage across all subject areas and geographical regions, but still lacks quality assurance and transparency about the resources covered, and cannot serve as a sole source for research performance evaluation (Halevi, 2017).
Non-source titles and publication were identified and quantified for each level of the Teaching area Qualis classification. We verified now many of the non-source titles met the following requirements for data search and recovery, which are common to bibliographic databases:

- Article title, abstract and keywords in English
- A DOI assigned to each article
- Use of ORCID for author identification (ORCID, 2019)

The number of articles reported by the Teaching area programs published in the non-source titles was counted for each Qualis classification level, to determine the number of non-source publications. Though knowing that the coverage of indexed titles in any database is subject to gaps, we did not search for individual articles amongst the indexed titles to confirm if indeed they are to be found in WoS and Scopus, as this falls out of the adopted definition for non-source publications.

Employing the free online tool WordArt available at https://wordart.com/, a word cloud was created using the title of non-source journals, to illustrate the concentration of publications per title. Another two word clouds were created from article titles from: a) non-source publications and b) indexed publications. For each, a table with the most frequent words was extracted and word variants have been manually categorized, as for example: “Ensino”, “Ensinar”, “Ensinando”, “Teach”, “Teaching”, “Enseñanza”. New clouds were created with the aggregated word categories, and the 10 most frequent ones were compared.

Results and Discussion

From the CAPES Indicator Spreadsheet for the TEACHING area (Output List tab) we´ve recovered 14,038 records of articles published in 2,320 peer-reviewed journals in the period between 2013 and 2016. After removing the duplicate records (19.7%) we found 11,270 unique articles published in 2,267 journals.
Analysis showed that the articles published by the Teaching area were very concentrated in a small number of journals. Only 209 Journals (9%) published 10 or more articles and where considered for further analysis. These titles published 6,697 (59.4%) of all articles in the period 2013-2016, were as almost half of the titles published only one article from the Teaching area, as can be seen in Table 4.

Table 4. Distribution of articles per title for the Teaching area.

Published articles from Teaching area	*Journals*	*Articles*
10 or more	209 (9.0%)	6,697 (59.4%)
2 - 9	956 (42.2%)	3,445 (30.6%)
1	1,102 (48.6%)	1,102 (9.8%)

This concentration of publications in a selected number of titles can be observed in the word cloud depicted in Figure 1.

Figure 1. Word clouds using the titles from non-source journals used by the Brazilian graduation programs in the Teaching area between 2013-2016.

We proceeded to verify the coverage of the 209 relevant journals for the Teaching area and found that Latindex is the database with the highest coverage, with 144 (68.9%) of titles, followed by Google Scholar with 135 (64.6%), DOAJ with 80 (38.3%), Scopus with 36 titles (17.2%), WoS with 34 titles (16.3%), SciELO with 31 titles (14.8%) and ERIC with 3 titles (1.4%), as illustrated in Table 5.

Table 5. Coverage of journal titles in databases and search engines used in the Teaching area´s criteria for journal classification, per Qualis Classification Level (QCL).

Titles QCL	*Latindex*	*Google Scholar*	*DOAJ*	*Scopus*	*WoS*	*SciELO*	*ERIC*
ALL =209	**68.9% (n=144)**	**64.6% (n=135)**	**38.3% (n=80)**	**17.2% (n=36)**	**16.3% (n=34)**	**14.8% (n=31)**	**1.4% (n=3)**
A1=20	85% (17)	95% (19)	70% (14)	70.0% (14)	35.0% (7)	75.0% (15)	15% (3)
A2=43	88.4% (38)	90.7% (39)	48.8% (21)	7.0% (3)	11.6% (5)	11.6% (5)	0
B1=48	83.3% (40)	56.3% (27)	41.7% (20)	2.1% (1)	16.7% (8)	6.3% (3)	0
B2=34	58.8% (20)	52.9% (18)	20.6% (7)	20.6% (7)	14.7% (5)	2.9% (1)	0
B3=38	39.5% (15)	55.3% (21)	28.9% (11)	23.7% (9)	18.4% (7)	13.2% (5)	0
B4=14	57.1% (8)	42.9% (6)	21.4% (3)	7.1% (1)	0% (0)	7.1% (1)	0
B5=12	50% (6)	41.7% (5)	33.3% (4)	8.3% (1)	16.7% (2)	8.3% (1)	0

Scoring: A1 = 100, A2 = 85, B1 = 70, B2 = 55, B3 = 40, B4 = 25, B5 =10.

We found that 162 (77.5%) titles are not indexed in WoS or Scopus and, as we previously defined, were considered non-source journals. As expected, due to the Qualis criteria adopted by the Teaching area, the percentage of non-source journals was relatively low (25%) for titles classified as A1, but exceeded 76% for all classification levels between A2 and B5. In the period of 2013-2016 these titles published 5,255 (78.5%) non-source articles, which cannot be found in WoS nor in Scopus. Table 6 contains the total number of titles and publications in each Qualis classification level (QCL) for the Teaching area, as well as the number of non-source titles and publications.

Table 6. Non-source titles and non-source articles published by the Teaching Area between 2013-2016, per Qualis Classification Level (QCL).

QCL	*Total titles*	*Non-Source Titles*	*Total Publications*	*Non-Source Publications*
ALL	**209**	**77.5% (n=162)**	**6,697**	**78.5% (n=5,255)**
A1	20	25% (5)	1,030	19.1% (197)
A2	43	88.4% (38)	2,044	95.9% (1961)
B1	48	83.3% (40)	1,483	89.8% (1331)

B2	34	79.4% (27)	927	79.3% (735)
B3	38	76.3% (29)	707	82% (580)
B4	14	92.9% (13)	254	95.7% (243)
B5	12	83.8% (10)	252	82.5% (208)

Non-source titles = journals not covered by WoS Core Collection or Scopus
Non-source publications = articles published in non-source journals

Analysing the non-source titles, we found out that only 12 (7%) journals have article titles, abstracts and keywords (TAK) in English, a DOI assigned to every publication and require the author´s ORCID to submit an article. 78% of non-source titles had TAK in English but only 43% assigned a DOI to their publications. The author´s ORCID is only mandatory for submitting an article in 9% of the non-source titles, but optional in 56%, as can be observed in Figure 1, which contains the percentage of non-source titles which have TAK in English, DOI for publications and ORCID for authors.

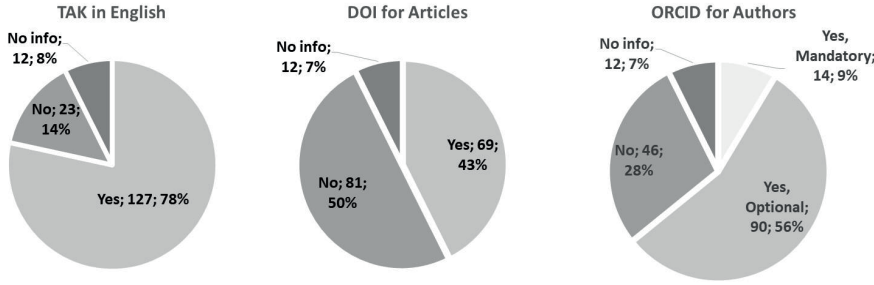


Figure 2. Percentage of non-source titles with Article Title, Abstracts and Keywords (TAK) in English, DOI assigned to publications and ORCID for authors.

A word cloud was created for the article titles from non-source publications and from indexed publications to illustrate the most frequent word categories used by the Teaching area between 2013 and 2016. In both cases “Ensino” (Teaching), “Professor” (Teacher) and “Educação” “Education” are within the most frequent word categories, as can be observed in Figure 3.

Figure 3. Word clouds using the article titles from non-source (left) and indexed publications (right), published by the Brazilian graduation programs in the Teaching area 2013-2016.

There is no significant difference between the most frequent word categories used in the non-source and indexed publication’s titles, except for their position in a rank of occurrence. Of special interest is the word “Brasil” and its variants, which appear in the 5th position, with 159 (11%) of article titles among the indexed publications. Within the non-source publications, “Brasil” and its variants appear in 304 (5.7%) of the article titles and are the 12th most frequent

word category. A more robust analysis would be necessary to understand if the articles that contain "Brasil" or variants in their title tend to be more focused on topics of local interest. This would require access to the article abstracts, key words and references, which are not available through the Sucupira Platform. Table 7 contains the 10 most frequent words for non-source and indexed publications.

Table 7. Most frequent word cartegories in the titles of articles published by the Brazilian graduation programs in the Teaching area between 2013-2016.

Non-source publications (n=5,255)			*Indexed publications (n=1,442)*	
Frequency	**Word**	**Rank**	**Word**	**Frequency**
1,899	Ensino (Teaching)	1	Ensino (Teaching)	373
1,121	Matemática (Mathematics)	2	Professor (Teacher)	268
1,108	Educação (Education)	3	Educação (Education)	253
1,096	Professor (Teacher)	4	Ciência (Science)	233
995	Ciência (Science)	5	Brasil (Brazil)	159
639	Formação (Training)	6	Formação (Training)	140
546	Aprendizagem (Learning)	7	Química (Chemistry)	140
479	Aluno (Student)	8	Matemática (Mathematics)	127
457	Escola (School)	9	Saúde (Health)	106
390	Didática (Didactics)	10	Análise (Analysis)	103

Non-source publications = articles published in non-source journals
Indexed publications = articles published in journals covered by WoS or Scopus

Conclusion

The Teaching area in Brazil publishes 77.5% of its research output in journals not covered by WoS or Scopus, confirming our assumption that any assessment of the Brazilian research in Teaching or Science Education should not rely solely in WoS or Scopus as sources of data. The comparison of word clouds produced with article titles of non-source publications did not show significant differences from the indexed publications, to allow us to distinguish any local or regional specific focus of research, which could be one of the reasons for not being covered by large international databases. The lack of a DOI assigned to each article in half of the non-source journals represents an obstacle to obtaining the articles´ meta data and full-text for more detailed analysis. The fact that most journals do not require an universal identifier for authors demands additional efforts in bibliometric studies to identify and disambiguate authors.

These characteristics are reflected in the data contained in the Sucupira platform and impact its potential to become the main source of data for bibliometric studies about the Brazilian scientific production. Nevertheless, its unique breadth of journals and publications makes it an essential source, in addition to international bibliometric databases, to obtain a comprehensive account of Brazil´s research activity and output in the Humanities, Social Sciences, or multidisciplinary areas.

References

Boyack, K. K. (2014). Including cited non-source items in a large-scale map of science: What difference does it make? *Journal of Informetrics*, 569-580.

Butler, L. V. (2006). Extending citation analysis to non-source items. *Scientometrics, 66*(2), 327-343.

CAPES. (2014). *CAPES Notícias*. Fonte: Coordenação de Aperfeiçoamento de Pessoal de Nível Superior: http://www.capes.gov.br/36-noticias/6810-capes-lanca-plataforma-sucupira-para-gestao-da-pos-graduacao

CAPES. (2017a). *Documento de Área Ensino*. Brasilia. Acesso em 19 de 12 de 2017, disponível em http://www.capes.gov.br/images/documentos/Documentos_de_area_2017/DOCUMENTO_AREA_ENSINO_24_MAIO.pdf

CAPES. (2017b). *Planilha de Indicadores*. Retrieved from Plataforma Sucupira: http://avaliacaoquadrienal.capes.gov.br/home/planilhas-de-indicadores

CAPES. (2017c). *Qualis*. Fonte: Coordenação de Aperfeiçoamento de Pessoal de Nível Superior: http://www.capes.gov.br/acessoainformacao/perguntas-frequentes/avaliacao-da-pos-graduacao/7422-qualis

CAPES. (2017d). *Sobre a Avaliação*. Fonte: Coordenação de Aperfeiçoamento de Pessoal de Nível Superior: http://www.capes.gov.br/avaliacao/sobre-a-avaliacao

CAPES. (2018). *GEOCAPES*. Fonte: Coordenação de Aperfeiçoamento de Pessoal de Nível Superior: https://geocapes.capes.gov.br/geocapes/

CAPES. (2019). *Quem participa*. Fonte: Portal de Periódicos: http://www-periodicos-capes-gov-br.ez45.periodicos.capes.gov.br/index.php?option=com_pcontent&view=pcontent&alias=quem-participa&Itemid=110

CHC. (2019). *Sobre a CHC*. Fonte: Ciência Hoje das Crianças: http://chc.org.br/sobre-a-chc/

Chi, P.-S. (2013). Do non-source items make a difference in the Social Sciences? *14th International Society of Scientometrics and Informetrics Conference*, (pp. 612-625). Viena.

Clarivate Analytics. (2018, 12 5). *Master Journal List*. Retrieved from Master Journal List: http://mjl.clarivate.com/

Clarivate Analytics. (2019). *The Impact Factor*. Retrieved from https://clarivate.com/essays/impact-factor/

CNPq. (2016). *Painel Lattes - mapa estatístico*. Fonte: Conselho Nacional de Desenvolvimento Científico e Tecnológico : http://estatico.cnpq.br/painelLattes/mapa/

CNPq. (2019). *Sobre a Plataforma Lattes*. Fonte: Conselho Nacional de Desenvolvimento Científico e Tecnológico: http://lattes.cnpq.br/

Collazo-Reyes, F. (2013). Growth of the number of indexed journals of Latin America and the Carribean: the effect on the impact of each country. *Scientometrics, 98*, 197-209. Retrieved 10 17, 2018

Coutinho, R. X. (2012). Brazilian scientific production in science education. *Scientometrics, 92(3)*, 697-710. Retrieved 3 5, 2013

Dagiene, E. S. (2015). Dynamics between National Assessment Policy and Domestic Academic Journals. *15th International Conference of the International Society for Scientometrics and Informetrics*, (pp. 1191-1193). Istanbul.

De Meis, L. A. (2007). The impact of science in Brazil. *IUBMB Life, 59(4-5)*, 227-234. Retrieved 3 5, 2013

DOAJ. (2019). *About DOAJ*. Retrieved from Directory of Open Access Journals: https://doaj.org/

Elsevier. (2018). *Scopus Content*. Retrieved from Elsevier Solutions: https://www.elsevier.com/solutions/scopus/how-scopus-works/content

ERIC. (2019). *About ERIC*. Retrieved from Institute of Education Sciences: https://eric.ed.gov/

Google Scholar. (2019). *About Google Scholar*. Retrieved from About Google Scholar: https://scholar.google.com/intl/en/scholar/about.html

Halevi, G. M.-I. (2017). Suitability of Google Scholar as a source of scientific inforation and as a source of data for scientific evaluation - Review of the Literature. *Journal of Informetrics*.

Hicks, D. (1999). The difficulty of achieving full coverage of international social science literature and the bibliometric consequences. *Scientometrics*, 193-215.

Hicks, D. (2012). Performance-based university research funding systems. *Research Policy*, 251-261.

Hicks, D. W. (2010). Coverage and Overlap of the New Social Sciences and Humanities Journal Lists. *Journal of the American Society for Information Science and Technology*, 284-294.

Latindex. (2019). Retrieved from Online Regional Information Systema for Scientific Journals from Latin America, the Caribbean, Spain and Portugal: http://www.latindex.org/latindex/inicio?lang=en

Leta, J. G. (2006). Science in Brazil. Part 2: sectoral and institutional research profiles. *Scientometrics, 67*(1), 87-105.

Madrid-Martín, M. J.-F.-M. (2017). Revistas brasileñas de Educación en Scopus: un análisis bibliométrico. *Biblios, 67*. Retrieved from http://biblios.pitt.edu/ojs/index.php/biblios/article/view/344

Medina-Jerez, W. (2018). Science Education Research Trends in Latin America. *International Journal of Science and Mathematics Education, 16*, 465-485.

Moed, H. (2005). Citation Analysis in Research Evaluation. Dodrecht: Springer.

ORCID. (2019). *What is ORCID*. Retrieved from ORCID: https://orcid.org/about

RBECT. (2019). *Focus and Scope*. Fonte: Revista Brasileira de Ensino de Ciência e Tecnologia: https://periodicos.utfpr.edu.br/rbect/about/editorialPolicies

SciELO . (2018). *Jornals by publication years*. Retrieved from Collection Indicators: http://www.scielo.org/php/level.php?lang=en&component=42&item=26

UFRGS. (2018). *Universidade Federal do Rio Grande Do Sul*. Fonte: Programa de Pós-Graduação em Ensino de Física: http://www.if.ufrgs.br/ppgenfis/index.php

Which courses to follow? On the relationship between the mobility of China-connected scholars and their academic performance

Zhenyue Zhao[1], Lele Kang[2], Chao Min[3], Yi Bu[4], Yiyang Bian[5] and Jiang Li[6]

[1] *njujack@163.com,* [2] *lelekang@nju.edu.cn,* [3] *mc@nju.edu.cn,* [5] *bianyiyang@nju.edu.cn,* [6] *lijiang@nju.edu.cn*
School of Information Management, Nanjing University, China

[4] *buyi@iu.edu*
Center for Complex Networks and Systems Research, School of Informatics, Computing, and Engineering, Indiana University, Bloomington, Indiana, U.S.A.

Abstract

A large number of oversea native elites were brought back to China by policy in the past decade; however, a handful of them moved overseas again. Taking advantages of the ORCID website, this study investigates 2,425 China-connected scholars' academic performance in terms of their migration. The results show that (1) the outflow-inflow of scholars between China and other countries was basically symmetrical except for several countries such as Pakistan and Japan; (2) The trends show that scientists published more academic articles; and (3) Scientists are more likely to become corresponding authors after they move to China.

Introduction

The migration of scientific researchers from developing countries to developed countries is often referred to as 'brain drain' (Beine, Docquier, & Oden-Defoort, 2011). The 'brain drain' occurs because of the better living conditions and more opportunities for researchers to distinguish themselves in their careers. While the 'brain drain' was still haunting most developing countries, 'brain gain' appeared in some developing countries which significantly rewarded the return of oversea scientists (Beine et al., 2011), i.e., some gained brains in the developed countries start moving back to their home countries. It is also called 'reverse brain drain' (or 'forth-back'). Relevant studies explored whether the gained brain really contributed to the development of the destination countries (Saxenian, 2002), or whether they helped to construct a better scientific community (Dyachenko, 2017). Recently, scientists are found to move forth-and-back in developing and developed countries (Daugeliene & Marcinkeviciene, 2015), making 'brain circulation' a new phenomenon in developing countries.

After being a major brain sending country for decades in the last century, China has released a series of policies intending to bring back native elites from abroad and recruiting foreign experts from other countries (Jonkers, 2008). As reported by Ministry of Education of the People's Republic of China, the number of oversea learners having returned to China after receiving their degrees reached 480,900 in 2017 (MOE, 2017). However, there is another side of the coin. Dramatically, in recent years, a tendency has emerged that Chinese scholars who came back from abroad returned overseas, i.e. 'forth-back-forth'. Examples include an outstanding professor of life sciences, who joined Tsinghua University after graduating from Princeton University in 2007 and returned to Princeton in 2017. Zweig (2006) pointed out that the weakness of Chinese academic system is the major reason why scholars move forth-back-forth. It is debatable whether the productivity of the returned oversea scholars is threatened by the behindhand academic system. To address this question, this study compares the research performance of scientists when they were in China and when they were not, in a 10-year timespan from 2008 to 2017.

Related work

Existing literature suggested that transnational mobility rates of scientists are larger than average migration rates of the whole society (Børing et al., 2015; Czaika & Orazbayev, 2018; Ioannidis, 2004;), since scientists are led by the desires both for better living conditions and for a glorious career (Geuna & Shibayama, 2015; Laudel, 2005; Jöns, 2007). Recent studies show that a global pattern of 'brain circulation' has formed, abandoning old patterns measuring the net gains and losses (Sugimoto et al., 2017).

Studies on the performance of the mobile scholars revealed that scholars with high mobility generally perform better in scientific research than non- or less-mobile ones (Sugimoto et al., 2017). Besides, the high-skilled from abroad (including scientists, engineers, etc.) indeed contribute to the development of the destination countries (Saxenian, 2002). These studies account for the popularity of scientific mobility among scholars and the willingness of countries to recruit overseas scholars (Appelt et al., 2015; Azoulay, Ganguli, & Zivin, 2017).

China's current situation stands out featuring a high returning rate of native scholars working abroad. The returning of scholars, contributing a great part of the brain circulation, is another hot topic recently. Several studies on this phenomenon discussed the chemical reactions between the returned scholars and the academic environments. This included the research performance (Jonkers & Cruz-Castro, 2013), the collaboration behavior (Jonkers & Cruz-Castro, 2013) of the scholar and knowledge diffusion (Kahn & MacGarvie, 2016), and the amelioration of academic ecologies in the receiving institutes or countries (Li, Miao & Yang, 2015). Among those, the research performance of scholars that returned to China is the main course under discussion in this paper.

Methodology

This study collected data of China-connected scholars (whose doctoral education and work experience include but are not limited to Mainland China) from www.orcid.org, including their academic resume. We obtained a sample dataset comprised of 2,425 scholars by applying the following two criteria. First is that he/she should be a doctoral student or a researcher/professor or both, affiliated to at least two countries/territories between 2008 and 2017, one of which must be Mainland China. Second is that he/she should have published at least one paper indexed by the Web of Science database (SCI-E, SSCI, A&HCI) by the end of 2018.

The ORCID, i.e., Open Researcher and Contributor ID, is a non-proprietary alphanumeric code to uniquely identify authors and contributors in scientific literature. Name disambiguation in science literature is hence avoided by using ORCIDs of authors. On www.orcid.org, there are ~0.75 million scientists from all over the world, and their academic experience dates back to the year of 1913. We identified 3,513 scholars whose affiliations meet the first criterion. Next, by using the ORCIDs, we searched for the publications of the 3,513 scholars in the WoS database (SCI-E, SSCI, A&HCI). We found that 1,088 scholars did not have a publication. The remaining 2,425 scholars have published 61,862 papers, of which 49,321 (79.7%) were published between 2008 and 2017.

Among the 2425 scholars, we roughly identified about 2100 Chinese (at least ethnically) by names. Each of the 2,425 scholars' affiliations in publications was extracted from the field 'C1' in the WoS. Then, we matched the information with his/her affiliations on the ORCID website to determine his/her outputs for each affiliation. We used machine processing first and successfully matched the WoS affiliations to corresponding ORCID affiliations for 36,582 papers. After manual inspection, the number increased.

The mobility phase of each scientist between 2008 and 2017 was divided into at most three phases as show in Figure 1: P_p (the period pre-returning to China), standing for the period that he/she spent abroad before coming to Mainland China; P_r (the period returning to China), referring to the period he/she spent in Mainland China; and P_a (the period after leaving China), indicating the period he/she spent after leaving Mainland China. Two of the three periods were available for most of the 2,425 scholars, i.e., P_p and P_r, or P_r and P_a. If a scholar stayed in China more than once, the last two periods repeat like P_p, P_r, P_a, P_r, and P_a.

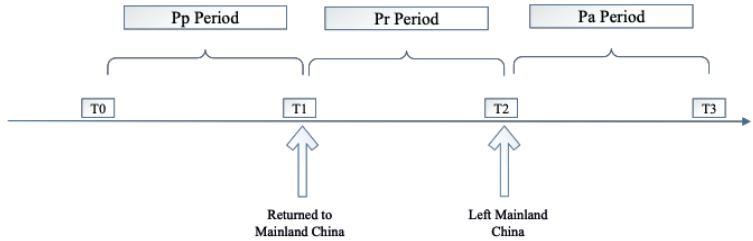


Figure 1. An illustration on the three phases of migration

If more than one affiliation was simultaneously used by a single author in an article, we only considered the latest affiliation matched with the ORCID data. We ignored an author's affiliation(s) which appeared in his/her articles but did not appear in any of the three periods. Paired samples t-test and Chi-squared test were applied to determine the differences of scholars' outputs between P_p and P_r, and between P_r and P_a. Two performance indicators were used: average number of papers published per month and the proportion of corresponding-author papers.

Primary results

Table 1 displays the descriptive statistics of our dataset extracted from the ORCID website and the WoS database.

Table 1. Descriptive statistics of the dataset

Field name	*N*	*Min.*	*Max.*	*Avg.*	*Description*
ORCID	2,425	-	-	-	Containing 2,425 unique ORCIDs
Mobility frequency	2,425	1	45	1.31	Times of mobility of a scholar during 2008-2017
Duration_ P_p	1,436	1	105	41.3	Months of period P_p
Duration_ P_r	2,373	1	197	47.7	Months of period P_r
Duration_ P_a	1,101	1	517	43.9	Months of period P_a
Papers_ P_p	1,436	0	374	5.78	Number of papers published in P_p
Papers_ P_r	2,373	0	197	9.74	Number of papers published in P_r
Papers_ P_p	1,101	0	107	6.25	Number of papers published in P_a
Proportion_ P_p	1,436	0	100%	15.9%	Proportion of corresponding author papers during P_p
Proportion_ P_r	2,373	0	100%	26.9%	Proportion of corresponding author papers during P_r
Proportion_ P_a	1,101	0	100%	27.2%	Proportion of corresponding author papers during P_a

Symmetrical inflow-outflow relationship between China and other countries/territories

The flowing pattern of the selected scholars during 2008-2017 is shown in Figure 2. It is clear to see that the U.S. has a significant impact over China in both 'brain gain' and 'brain drain' during the 10 years. It was not only the country where most scholars left China for (36.8%, or

425 scholars), but also the country who sent the most scientists to China (37.8%, or 614 scholars).

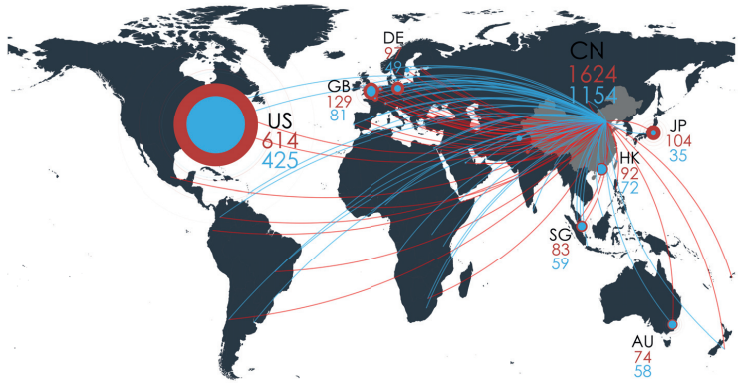


Figure 2. Flowing pattern of scholars during 2008-2017. Note: Red curves indicate scientists' returning to China from abroad and blue curves mean scientists flowing from China to abroad; the size of circles represents the number of moving scientists.

We found a symmetrical inflow-outflow relationship between China and most other countries/territories. This indicates that China and other countries/territories both lost and gained scholars during 2008-2017, but neither suffered from a sharp drain or gain. This phenomenon, if we take a broader view on the mobility of scholars globally, could be a positive boost for 'brain circulation'. We also found countries featuring asymmetrical relationships. Pakistan has gained four times the number of scholars it has drained to China, while China has gained almost four times the number of scholars it has drained to Japan. Overall, China has gained more (1,624 scholars) than has drained (1,154 scholars) during the decade.

Scientists gained higher productivity and more chances of being corresponding author after moving to China

Before matching the papers to the exact affiliations with which they were produced, we drew two lines (see Figure 3) for the yearly output of 187 scholars whose working experiences covered 4 years/5 years before/after they moved to China. After moving to China, the 187 scholars were observed to have gained more rapid growth both in productivity and in the proportion of corresponded papers.

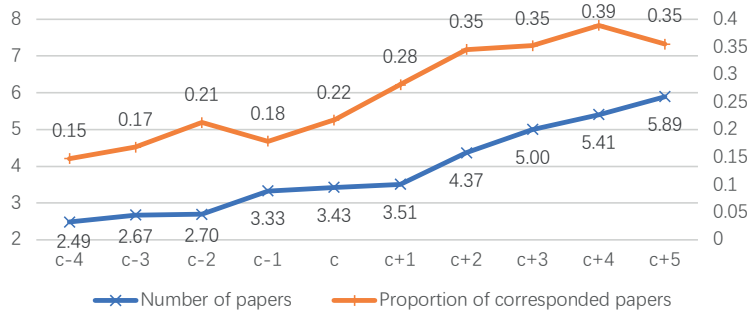


Figure 3. Scientists' academic performance before and after they moved to China

Then we obtained 1,389 pairs of P_p - P_r working experiences for 1,366 scientists, and 966 P_r - P_a pairs for 956 scientists. Note that one scientist has two paired data if he/she returned to China twice. Paired samples t-test was applied to examine scientists' research performance difference between P_p and P_r, and between P_r and P_a. After returning to China, scientists' monthly publications significantly increased from 0.126 to 0.212 on average ($p<0.01$, paired-samples t-test), whereas the number significantly decreased from 0.162 to 0.146 on average ($p<0.05$, paired-samples t-test) when they left China for other countries/territories. Besides, scholars who returned to China from overseas published significantly more corresponding-authored articles. Their proportions of corresponding-authored papers increased significantly averagely from 15.7% to 36.5% ($p <0.01$, paired-samples t-test). In comparison, the proportion of corresponding-authored papers for those who left China only slightly increased from 13.8% to 18.3% ($p <0.01$, paired-samples t-test). It is also noticeable that 572 scientists (41.2% of the total) gained first experience of being a corresponding author in China, while the number is 198 (20.5%) for those who left China for other countries/territories ($p<0.01$, Chi-squared test).

As we see this proportions of being corresponding authros generally increase after scholars came to China, we infer that China has facilitated the construction of their research teams and laboratories for those moving to China, and this may also partly account for the rapid growth in their productivity. Another reason for the productivity growth is the extreme competitive environment and the quantity-based evaluation system in China. The scholars, especially who received generous investment from the governments, had to keep publishing in order not to be left behind or even eliminated (Chen & Shu, 2017).

Limitations and future work

There are inevitable limitations in this study. It can be inferred that scientists who are more willing to be involved in international academia are more likely to fill in their information on the ORCID website, leading to the bias of sampled scholars. In addition, the loss of some authors' ORCIDs in WoS resulted in the incomplete match of their affiliations in ORCID and WoS data.

We will scrutinize the influential indicators such as scholars' academic age, disciplinary background, co-authorship and citation impact and investigate their impacts on mobility. We also expect our future results could reveal the motivation of their moving.

Acknowledgements

The authors acknowledge the National Natural Science Foundation of China (NSFC Grant No. 71673242) and Jiangsu Province Social Sciences Foundation (Grant No. 18TQC005) for financial support. The authors are grateful to Ms. Siyu Hang, Ms. Zhaonan Guo, Mr. Ziwei Han and Ms. Xinyi Guan for double-checking the validity and availability of the data used in our research.

References

Appelt, S., Beuzekom, B. V., Galindo-Rueda, F., & Pinho, R. D. (2015). Which factors influence the international mobility of research scientists?. *OECD Science Technology & Industry Working Papers.*

Azoulay, P., Ganguli, I., & Zivin, J. G. (2017). The mobility of elite life scientists: Professional and personal determinants. *Research policy, 46*(3), 573-590.

Beine, M., Docquier, F., & Oden-Defoort, C. (2011). A panel data analysis of the brain gain. *World Development, 39*(4), 523-532.
Børing, P., Flanagan, K., Gagliardi, D., Kaloudis, A., & Karakasidou, A. (2015). International mobility: Findings from a survey of researchers in the EU. *Science and Public Policy, 42*(6), 811-826.
Czaika, M., & Orazbayev, S. (2018). The globalisation of scientific mobility, 1970–2014. *Applied Geography*, 96, 1-10.
Daugeliene, R., & Marcinkeviciene, R. (2015). Brain circulation: theoretical considerations. *Inzinerine Ekonomika Engineering Economics, 33*(3), 49-57.
Dyachenko, E. L. (2017). Internal migration of scientists in Russia and the USA: the case of physicists. *Scientometrics, 113*(1), 105-122.
Gaulé, P. (2014). Who comes back and when? Return migration decisions of academic scientists. *Economics Letters, 124*(3), 461-464.
Geuna, A., & Shibayama, S. (2015). Moving out of academic research: why do scientists stop doing research?. In *Global mobility of research scientists* (pp. 271-303).
Ioannidis, J. P. (2004). Global estimates of high-level brain drain and deficit. *The FASEB Journal, 18*(9), 936-939.
Jonkers, K. (2008). *A comparative study of return migration policies targeting the highly skilled in four major sending countries*. Retrieved January 31, 2019 from: http://cadmus.eui.eu/bitstream/handle/1814/9454/MIREM%20AR2008_05.pdf?sequence=1&isAllowed=y
Jonkers, K., & Cruz-Castro, L. (2013). Research upon return: The effect of international mobility on scientific ties, production and impact. *Research Policy, 42*(8), 1366-1377.
Jöns, H. (2007). Transnational mobility and the spaces of knowledge production: a comparison of different academic fields. *Social Geography Discussions, 2*(50), 79-119.
Kahn, S., & MacGarvie, M. (2016). Do return requirements increase international knowledge diffusion? Evidence from the Fulbright program. *Research Policy, 45*(6), 1304-1322.
Laudel, G. (2005). Migration currents among the scientific elite. Minerva, 43(4), 377-395.
Li, F., Miao, Y., & Yang, C. (2015). How do alumni faculty behave in research collaboration? An analysis of Chang Jiang Scholars in China. *Research Policy, 44*(2), 438-450.
MOE. (2018). 2017 *sees increase in number of Chinese students studying abroad and returning after overseas studies*. Retrieved January 31, 2019 from: http://en.moe.gov.cn/News/Top_News/201804/t20180404_332354.html
Quan, W., Chen, B., & Shu, F. (2017). Publish or impoverish: An investigation of the monetary reward system of science in China (1999-2016). *Aslib Journal of Information Management*, 69(5), 486-502.
Saxenian, A. (2002). Brain circulation: How high-skill immigration makes everyone better off. *Brookings Review, 20*(1), 28-31.
Sugimoto, C. R., Robinson-García, N., Murray, D. S., Yegros-Yegros, A., Costas, R., & Larivière, V. (2017). Scientists have most impact when they're free to move. *Nature News, 550*(7674), 29.
Zweig, D. (2006). Competing for talent: china's strategies to reverse the brain drain. *International Labour Review, 145*(1-2), 65-90.

Comparison of classification-related differences in the distribution of journal articles across academic disciplines: the case of social sciences and humanities in Flanders and Norway (2006-2015)

Linda Sīle[1], Raf Guns[1], Frédéric Vandermoere[2], and Tim C. E. Engels[1]

[1]*Linda.Sile@uantwerpen.be, Raf.Guns@uantwerpen.be, Tim.Engels@uantwerpen.be*
University of Antwerp, Faculty of Social Sciences, Centre for R&D Monitoring (ECOOM), Middelheimlaan 1, 2020 Antwerp (Belgium)

[2]*Frederic.Vandermoere@uantwerpen.be*
University of Antwerp, Faculty of Social Sciences, Department of Sociology, Sint-Jacobstraat 2-4, 2000 Antwerp (Belgium)

Abstract

Even though bibliometric analyses often rely on different disciplinary classifications, it is not known to what extent the choice of classification influences bibliometric findings. Here we explore differences in the distribution of articles across disciplines using four different classifications. We use data on social sciences and humanities (SSH) journal articles from comprehensive bibliographic databases in Belgium (Flanders) and Norway (2006-2015). In our analysis we use the original classifications used in VABB-SHW and Cristin, the Flemish and the Norwegian databases, Web of Science subject categories, and Science-Metrix journal classification.

Preliminary findings show that different classifications lead to considerable differences in the total number of SSH journal articles. For example, the percentage difference in the number of SSH publications for Norway is 17,5% when comparing the Science-Metrix and the original classification. This implies that there is a substantial number of publications in disciplinary terms residing on the boundaries between SSH and other knowledge domains. In contrast, on discipline level the differences due to the classification are small (the average difference in share is less than 2 p.p.). This might mean that if one employs a scheme with relatively broad categories then the choice of classification is of minor importance.

Introduction

Over time, multiple methods have been developed to identify academic disciplines to which publication sets belong (for an overview see Gläser, Glänzel, & Scharnhorst, 2017). One can use content-based classification approach (Bensman, 2007), classify publications (or journals) on the basis of citations (Carley, Porter, Rafols, & Leydesdorff, 2017; Leydesdorff, Bornmann, & Zhou, 2016), use text-based algorithmic approaches (Eykens, Guns, & Engels, 2019) or a hybrid text and citation-based approach (Janssens, Zhang, Moor, & Glänzel, 2009). Finally, it is possible to use a more pragmatic approach and rely on already existing journal classifications (e.g. Leydesdorff & Rafols, 2009; Wang & Waltman, 2016).

Not all of these approaches, however, are suitable for bibliometric studies of research within the social sciences and humanities (SSH). While the citation-based approach is severely limited due to the low coverage of SSH in WoS and Scopus (Leydesdorff, Hammarfelt, & Salah, 2011), text-based approaches are challenged by the fact that SSH literature is scattered across multiple disciplinary databases (such as PhilPapers, ERIC, PsycNET). Knowledge of their comprehensiveness and comparability is limited. Furthermore, such specialised databases are not available for all disciplines within SSH. An alternative is to use data from the more comprehensive national databases (e.g. VABB-SHW in Belgium, Cristin in Norway).

National databases, although more suitable for bibliometric studies of SSH, often employ local classifications (Guns, Sīle, Eykens, Verleysen, & Engels, 2018; Kulczycki et al., 2018; Ossenblok, Engels, & Sivertsen, 2012). Considering the literature on classifications in general (e.g. Bowker & Star, 2000), we know that classifications carry traces from the contexts in which they originate. It is possible that in one context a journal is understood as belonging to a discipline X while in another it is perceived as belonging to a discipline Y. This aspect of

classifications has not been explored in relation to (local) disciplinary classifications for SSH journals. This, consequently, is a challenge since it is not known how valid is the use of such classifications for the calculation of bibliometric indicators in comparative settings. Accordingly, the purpose of this article is to explore how the choice of classification influences the distribution of articles across disciplines in SSH. To do so, we use data on SSH journal articles from Flanders (Belgium) and Norway (2006-2015) acquired from two national bibliographic databases. In our analysis we employ four different classifications: (a) the VABB-SHW cognitive classification (VABB-OECD), (b) the classification used in the Norwegian list of Scholarly Journals (NPU) (c) Web of Science subject categories (WOS-SC) and (d) Science-Metrix classification for journals (SM, Archambault, Beauchesne, & Caruso, 2011).
In what follows, we begin with a brief description of our data and methods. Second, we continue with preliminary findings. We present preliminary findings from our comparisons of the total number of SSH articles and the disciplinary structure for SSH. Finally, we draw links between our findings and implications for the use of bibliometrics in policy settings.

Data and methods

In this study we use data from two national bibliographic databases (VABB-SHW in Flanders and Cristin in Norway; for details on both databases see Sīle et al. 2018) and Web of Science (in-house database at ECOOM-Leuven, dataset retrieved on 23/07/2018). The analysis is conducted using eight datasets (A, B, C, D: four datasets for each country) delineated as follows. Data set A consists of peer-reviewed journal articles in SSH (2006-2015) by authors affiliated to universities (the 5 universities in Flanders and the 8 universities in Norway). The data sets B, C, and D are subsets of A (see Table 1). The data set B is limited to articles from the data set A that are indexed in the three main indices in WoS. The identification of WoS-indexed articles was carried out on article level using datasets retrieved the ECOOM-Leuven in-house WoS database and a string matching approach that allows for small differences in the matched references (for details see Sīle & Guns, 2019). The third data set C is limited to articles in journals which are included in a classification of journals developed by Science-Metrix (identified on the basis of ISSNs). The dataset D contains articles in journals that are both indexed in WoS and listed in the Science-Metrix classification.

Table 1 Samples and criteria used to delineate data subsets

Sample and criteria used to delineate data subsets	Dataset			
	A	B	C	D
Flanders	29648	12575	11542	9988
Norway	26007	9444	10399	7783
Articles indexed in Web of Science (SCIE, SSCI, AHCI)	-	Yes	-	Yes
Journals included in Science-Metrix list	-	-	Yes	Yes

To identify the extent which the choice of academic discipline classification influences bibliometric indicators for SSH, we, first assign each article to an academic discipline using four classifications. For comparability, all classifications are mapped to OECD FORD classification. Then we explore how the distribution of publications across disciplines changes depending on the classification. This exploration is carried out on the basis of percentage difference and arithmetic difference in share comparisons. All analyses are carried using fractionalised counts at the author level. The use of fractionalised counts is more appropriate given our use of bibliometric indicators as proxies of research activities.
Percentage difference is acquired by, first, dividing the difference between the number of publications acquired using one classification (V1) and the number of publications acquired using another classification (V2) by the average of the value, and then multiplying by 100. This

equation is deemed more suitable for the analysis presented here since we do not prioritise and assume one of the classifications as correct (as it would be when calculating, for example, percentage error).

Also we point out a limitation of this study that is related to the comparative nature of the analysis. We use data from two different national databases. Even though both databases are assumed to be comprehensive databases for peer-reviewed scholarly publications (for Flanders, only for SSH), we are aware that there are some differences in databases setups that might alter our results. Given the focus of the study, we assume that the acquired level of accuracy is acceptable.

Preliminary findings

Preliminary findings of this study indicate that the choice of disciplinary classification for (SSH journals) has a modest influence on bibliometric representations of SSH research: the total number of articles in SSH is over- or underestimated. However, there is practically no influence on the disciplinary structure—the distribution of articles across disciplines. The greatest difference is in the total number of SSH publications rather than difference in share of articles in specific disciplines (Table 1).

We find, for example, that the percentage difference in the total number of SSH publications for Norway is 17,5% when comparing SM with the original classification. 1201,1 journal articles that are assumed to belong to SSH according to the NPU classification, are not assumed as such using SM classification. For Flanders, comparison of SM and the original classification reveals a percentage difference of 11,5%. Other comparisons reveal smaller differences. Percentage difference between WoS and the original classification is 3,5% and 8,5% respectively. Differences in the total number and the share of SS and H publications are minor and range from practically no difference (H, SM versus original classifications, both countries) to a difference of 6% (SS, WoS versus original classification, Norway). In all cases, differences are slightly greater for Norway.

Percentage differences on article level vary considerably by discipline (range: 1,5%-154%; *M*=31%; *SD*=35%; *Md*=18%). The percentage difference is especially high for the category 'Other social sciences' (comparing with WOS-SC: 154% for Flanders; 44% for Norway; comparing with SM: 143% for Flanders; 52% for Norway). Using the local classifications, the number in this category is lower. This might be an indication of a more conservative tendency in journal assignment to disciplinary categories: interdisciplinary research or research from new disciplines is perceived as belonging to one of the more established disciplines. This interpretation is supported also by the very low differences for categories 'Languages and literature' (comparing with WOS-SC: 2% for Flanders and Norway) and 'Educational sciences' (comparing with SM: 4% for Flanders and 5% for Norway). These low differences, however, are not consistent for the compared classifications (WOS-SC and SM) thus pointing out that also WOS-SC and SM carry assumptions on journal assignment to disciplinary categories that can be more or less in alignment to what is used in national contexts. These differences can certainly be partly be explained by the uneven distribution of articles across disciplines: percentage difference for categories with a low total number of articles (N<100) will appear more substantial than for categories with high number of articles (N>1000). Nevertheless, these findings indicate that there is more agreement on journal-discipline pairs for some disciplines (e.g. Languages and Literature, Arts) than others (e.g. Other social sciences, Social and economic geography, Media and communications) and the distribution of differences is not equivalent for both countries. These findings are in line with our theory-guided expectation that classifications carry traces from the contexts they are embedded in.

Overall, these findings suggest that even though differences due to classification can be observed in the total number of articles in SSH and in the total number by discipline, these differences do not substantially influence the disciplinary structure.

Discussion

Preliminary findings of our study show that the influence of the choice of disciplinary classification for journals in SSH is considerable when, firstly, delineating SSH publications and, secondly, when calculating research performance indicators based on absolute counts of publications per discipline. In contrast, when the focus is on research representations that are based on the relative number of articles by discipline (e.g. the disciplinary structure), the differences due to the choice of classification are minor.

These findings relate to those found in science mapping literature. For example, Rafols and Leydesdorff(2009) find substantial differences in journal classifications yet the structure of science maps that is acquired using these classifications is similar. It might be that the differences in percentage change and the absence of differences in the disciplinary structure we observe can be explained statistically (as in Leydesdorff & Rafols, 2009). The disciplinary structure is not only affected by the number of publications in each discipline, but also by the number of publications *in relation* to other disciplines. For major differences in the disciplinary structure, changes in the absolute number due to classification would need to be much greater than the ones observed here.

We found a difference when comparing the total number of SSH publications identified as such using different classifications. On the one hand, this means that a considerable number of SSH publications (and journals) are residing on the boundaries between SSH and other knowledge domains (e.g. Medical fields, Environmental sciences). This might have implications when larger knowledge domains are used in indicator construction or in the choice of evaluation approach (e.g. the case of Italy described by Ancaiani et al., 2015). On the other hand, the small differences in the disciplinary structure seems to suggest that the choice of academic discipline classification is of no importance since the results are altered only to a minor extent. However, two points can be highlighted. First, even though the differences we identify appear small (on average 2.2 p.p.), they can turn out crucial if linked to some reward mechanism (e.g. funding allocation or promotion). In such contexts, even a difference of 2 p.p. may have consequences especially for small yet highly specialised knowledge domains. Second, these small alterations might be a consequence of the mapping activity employed in this analysis. As noted, all the classifications that we used were mapped to OECD FORD to improve comparability. However, for SSH OECD FORD is limited only to 12 either established (e.g. Psychology) or broad disciplines (e.g. Media and Communications) and 2 residual categories (Other social sciences and Other humanities). Such structure carries a risk that research in more specific disciplines that have designated categories in some disciplines (e.g. History of Ideas or Science, Technology, and Society studies) is rendered invisible.

References

Ancaiani, A., Anfossi, A. F., Barbara, A., Benedetto, S., Blasi, B., Carletti, V., … Sileoni, S. (2015). Evaluating scientific research in Italy: The 2004–10 research evaluation exercise. *Research Evaluation*, *24*(3), 242–255. https://doi.org/10.1093/reseval/rvv008

Bensman, S. J. (2007). Garfield and the impact factor. *Annual Review of Information Science and Technology*, *41*(1), 93–155. https://doi.org/10.1002/aris.2007.1440410110

Bowker, G. C., & Star, S. L. (2000). *Sorting things out: classification and its consequences* (First paperback edition). Cambridge, Massachusetts London, England: The MIT Press.

Carley, S., Porter, A. L., Rafols, I., & Leydesdorff, L. (2017). Visualization of Disciplinary Profiles: Enhanced Science Overlay Maps. *Journal of Data and Information Science*, *2*(3), 68–111. https://doi.org/10.1515/jdis-2017-0015

Eykens, J., Guns, R., & Engels, T. C. E. (2019). Article Level Classification of Publications in Sociology: An Experimental Assessment of Supervised Machine Learning Approaches. *Proceedings of ISSI 2019*.

Gläser, J., Glänzel, W., & Scharnhorst, A. (2017). Same data—different results? Towards a comparative approach to the identification of thematic structures in science. *Scientometrics*, *111*(2), 981–998. https://doi.org/10.1007/s11192-017-2296-z

Guns, R., Sīle, L., Eykens, J., Verleysen, F. T., & Engels, T. C. E. (2018). A comparison of cognitive and organizational classification of publications in the social sciences and humanities. *Scientometrics*, *116*(2), 1093–1111. https://doi.org/10.1007/s11192-018-2775-x

Janssens, F., Zhang, L., Moor, B. D., & Glänzel, W. (2009). Hybrid clustering for validation and improvement of subject-classification schemes. *Information Processing & Management*, *45*(6), 683–702. https://doi.org/10.1016/j.ipm.2009.06.003

Kulczycki, E., Engels, T. C. E., Pölönen, J., Bruun, K., Dušková, M., Guns, R., … Zuccala, A. (2018). Publication patterns in the social sciences and humanities: evidence from eight European countries. *Scientometrics*. https://doi.org/10.1007/s11192-018-2711-0

Leydesdorff, L., Bornmann, L., & Zhou, P. (2016). Construction of a pragmatic base line for journal classifications and maps based on aggregated journal-journal citation relations. *Journal of Informetrics*, *10*(4), 902–918. https://doi.org/10.1016/j.joi.2016.07.008

Leydesdorff, L., Hammarfelt, B., & Salah, A. (2011). The structure of the Arts & Humanities Citation Index: A mapping on the basis of aggregated citations among 1,157 journals. *Journal of the American Society for Information Science and Technology*, *62*(12), 2414–2426. https://doi.org/10.1002/asi.21636

Leydesdorff, L., & Rafols, I. (2009). A global map of science based on the ISI subject categories. *Journal of the American Society for Information Science and Technology*, *60*(2), 348–362. https://doi.org/10.1002/asi.20967

Ossenblok, T. L. B., Engels, T. C. E., & Sivertsen, G. (2012). The representation of the social sciences and humanities in the Web of Science--a comparison of publication patterns and incentive structures in Flanders and Norway (2005-9). *Research Evaluation*, *21*(4), 280–290. https://doi.org/10.1093/reseval/rvs019

Rafols, I., & Leydesdorff, L. (2009). Content-based and algorithmic classifications of journals: Perspectives on the dynamics of scientific communication and indexer effects. *Journal of the American Society for Information Science and Technology*, *60*(9), 1823–1835. https://doi.org/10.1002/asi.21086

Sīle, L., & Guns, R. (2019). Article-level matching of Web of Science to a local database in a comparative context. *Proceedings of ISSI 2019*.

Wang, Q., & Waltman, L. (2016). Large-scale analysis of the accuracy of the journal classification systems of Web of Science and Scopus. *Journal of Informetrics*, *10*(2), 347–364. https://doi.org/10.1016/j.joi.2016.02.003

Table 2 Summary of differences in the total number of SSH, SS, and H journal articles for Flanders and Norway (2006-2015)

Value	Classification I	Classification II	Percentage difference, %			Difference in share, p.p.		
			Dataset B	Dataset C	Dataset D	Dataset B	Dataset C	Dataset D
The total number of SSH articles	VABB-OECD	WoS	3	-	4			
	NPU	WoS	8	-	9			
	VABB-OECD	SM	-	11	12			
	NPU	SM	-	17	18			
	WoS	SM	-	-	F: 8 N: 10			
The total number of SS articles	VABB-OECD	WoS	1	-	2	1,6	-	1,8
	NPU	WoS	6	-	7	1,6	-	1,9
	VABB-OECD	SM	-	14	14	-	2,7	1,9
	NPU	SM	-	20	21	-	2,7	1,8
	WoS	SM	-	-	F: 12 N: 14	-	-	F: 3,7 N: 3,6
The total number of H articles	VABB-OECD	WoS	3	-	7	1,9	-	2,2
	NPU	WoS	4	-	7	3,0	-	2,6
	VABB-OECD	SM	-	10	14	-	0,1	0,5
	NPU	SM	-	19	19	-	0,3	0,1
	WoS	SM	-	-	F: 20 N: 26	-	-	F: 2,6 N: 6,6

Table 3 Summary of differences in the disciplinary structure for Flanders and Norway (2006-2015) (F – Flanders, N – Norway)

Classification		Cosine similarity			Difference in share, p.p.								
					Mean			Median			Standard deviation		
I	II	Dataset B	Dataset C	Dataset D	Dataset B	Dataset C	Dataset D	Dataset B	Dataset C	Dataset D	Dataset B	Dataset C	Dataset D
VABB	WoS	0,99	-	0,99	-1	-	-1,2	-0,9	-	-1,2	1,3	-	1,5
NPU	WoS	0,98	-	0,98	-1,7	-	-1,8	-1	-	-1,1	2,2	-	2,5
VABB	SM	-	0,97	0,98	-	0,7	0,6	-	0,5	-0,1	-	2,6	2,6
NPU	SM	-	0,98	0,99	-	0,2	0,1	-	0,4	-0,1	-	1,5	1,3
SM	WoS	-	-	F: 0,98 N: 0,97	-	-	F: 1,7 N: 2,0	-	-	F: 1,5 N: 1,2	-	-	F: 2,1 N: 2,8

How Should We Measure Individual Researcher`s Performance Capacity Within and Between Universities – Social Sciences as an Example? A Multilevel Extension of the Bibliometric Quotient (BQ)

Rüdiger Mutz[1] and Hans-Dieter Daniel[2, 3]

[1] *mutz@gess.ethz.ch*
ETH Zurich, Professorship for Social Psychology and Research on Higher Education, Andreasstrasse 15, CH-8050 Zurich (Switzerland)

[2] *daniel@gess.ethz.ch*
ETH Zurich, Professorship for Social Psychology and Research on Higher Education, Andreasstrasse 15, CH-8050 Zurich (Switzerland)

[3] *hans-dieter.daniel@uzh.ch*
University of Zurich, Department of Psychology, Binzmuehlestrasse 14, CH-8050 Zurich (Switzerland)

Abstract

The assessment of individual research performance has become a major attraction for bibliometric researchers in recent years, and is dominated by the classic bibliometric indicator approach (e.g., h-index). Alternatively, a psychometric measurement approach is favored, which considers measurement errors. It is assumed that the "researcher's performance capacity" as a personal trait and competency is responsible for the individual research performance, which might vary randomly due to measurement errors. Five individual-level bibliometric variables served as items (e.g., number of articles in top 5%) to measure the competency. The central question of this contribution is how much variance in the "researcher`s performance capacity" is explained by differences between universities/subfields. With bibliometric data (Scopus) for a sample of 1,071 social scientists with Swiss university affiliations a one-dimensional scale ("Bibliometric Quotient", BQ) was created by means of a psychometric model, which has a high, but not perfect, reliability of r_{tt}=.84. The items were most suitable for scientists scoring above average. About 33% of the variance of the BQ is due to differences between the universities/subfields, and only 7% of the variance is due to differences between universities alone. A ranking only of Swiss universities in the social sciences does not necessarily make sense.

Introduction

The bibliometric-based measurement of individual research performance has attracted a great deal of attention in recent years, which is reflected in a multitude of literature on this topic (e.g., Abramo & D'Angelo, 2014; Bornmann & Marx, 2013, 2014; Bornmann & Mutz, 2011; Wildgaard, Schneider, & Larsen, 2014). "The evaluation of individual research performance is a fundamental tool for management, to inform decisions in areas such as faculty recruitment, career advancement, reward systems, grants awarding and projects funding." (Abramo, Cicero, & D`Angelo, 2013, p. 528). A large number of numerical indicators were developed at the level of the individuals. Wildgaard et al. (2014, p. 125) "reviewed 108 indicators that can potentially be used to measure performance on individual author-level". A prototype for such an indicator to assess individual research performance is the h-index.

The indicator approach, more or less adopted from economics, sociology and natural sciences, is less widespread in the sciences that deal with the individual, namely psychology or educational sciences. A major reason for this is the problem of random measurement errors, which are more significant at the level of individuals than at the level of institutions, and which are often not taken into account in the indicator approach (Abramo, D`Angelo, & Grilli, 2015; Karlsson et al., 2015). Due to different coverages of bibliographic databases, single publications

of individual researchers may be missing. Citation fluctuations might occur as a result of database updates (inclusion or removal of journals). Single highly cited publications do not reflect the overall work of a researcher. Such random fluctuations usually do not play a role at the institutional level, since they are averaged out during aggregation, especially if the size of institutions is high. Instead of relying on single indicators, psychology and educational sciences use a set of “indicators” called “items” that homogeneously measure a characteristic as theoretical construct that is not itself directly observable. These items have only a meaning within the construct they measure and may also be affected by measurement errors. A variety of psychometric test models have been developed to estimate quantitative test scores from empirical test data and thus measure a person's trait as a time-stable behavioral tendency. One characteristic, which has become generally known, is "intelligence": "A global concept that involves an individual's ability to act purposefully, think rationally, and deal effectively with the environment" (Wechsler, 1958, p. 7).

This measurement perspective motivated us to create a psychometric model based on bibliometric data to capture the scientific performance of researchers. With a modeling approach we hope to clarify questions of reliability, validity and fairness of the scale, and questions of dimensioning. These questions often remain unanswered in the classic indicator approach. A first model and a scale, the so-called Bibliometric Quotient, has already been developed and applied exemplarily to data from a sample of researchers in the field of social science methodology (Mutz & Daniel, 2018). Models have the advantage that they can be extended at will. Specific problems of a model can be solved by adding further model components in the hope the model fits the data better. In the indicator approach, special problems of an indicator (e.g., h-index) are often solved by the development of new indicators, whereby the letters of the alphabet are no longer sufficient to name the multitude of indicators (e.g., h-index, b-index, M-index), which has been developed.

This paper aims to extend the previous psychometric model of the Bibliometric Quotient (BQ) by a multilevel component, which considers differences between and within institutions of higher education. A topic, which attracts attention in the bibliometric indicator research, as well (Abramo, Cicero, & D`Angelo, 2012; Bonaccorsi & Cicero, 2016). How much variance in the BQ is explained by differences between universities? Institutional comparisons and rankings require a sufficient variability between institutions compared to the variability within institutions. The approach will be applied to bibliometric data on social scientists with Swiss university affiliations. The following research questions are in the focus:

1) Is it possible to create a one-dimensional scale form bibliometric data in order to measure the researcher's performance capacity? How reliable is the scale?
2) How high is BQ of social scientists from Switzerland?
3) How much variance in the BQ is explained by differences between universities? Is it possible to rank Swiss universities?
4) How strong are the relationships between the BQ and classic bibliometric indicators (e.g., h-index, total citations)?

Psychometric measurement model

Adopting the person-environment approach from psychology, we assume that the scientific performance of a researcher in the form of publications and their (citation) impact on the scientific field is based on the stable disposition or competency of a researcher (*person*), and on the research *environment*, in which he or she works (e.g., high citation level in life sciences). This competency is called “researcher`s performance capacity” (Harnad, 2008), as "competency of the researchers as authors to write influential papers" (Mutz & Daniel, 2018, p. 1284). To measure the theoretical construct, you need some individual specific variables,

called items, that repeatedly measure the same construct. In the case of measurement errors, it is expected that any kind of aggregation of items (i.e. scale) is more reliable than any single item. We assume that the "researcher's performance capacity" is the higher (in brackets the item labels),

- the higher the *scientific impact of the researcher's articles* in the researcher's scientific field is, measured as the number of publications that are in the top 5% in a scientific field (top5%, ITEM 1),
- the more that the publications have published as *first-author*, *mainly responsible for the article* (number of first author paper, ITEM 2),
- the higher the *impact of a single article*, the citation of the highest cited paper is (citation of the highest cited paper, ITEM 3),
- the more articles of a researcher have citations beyond the mean citation level of a field (MNCS, ITEM 4)
- the stronger the *short-term resonance* of the researcher in the scientific community is, measured as the total number of citations of the researcher's publications in a 5-year citation window (total citations in 5-year window, ITEM 5).

The *Item characteristic curve Poisson counts model* (ICCPCM) by Doebler, Doebler, and Holling (2014) serves as a psychometric model. It starts from a binary Rasch model as core model and add a frame model, which transforms the binary model to a Poisson count model.

Expressed in simple terms, the binary Rasch model, firstly introduced in bibliometrics by Alvarez and Pulgarin (1996a, 1996b, 1996c), assumes that the probability of an individual reaction to a binary item, for example, the probability that a researcher has published at least one paper in a relevant international journal or not, is a function of both the difficulty of the respective item, and the researcher`s competency (Andrich, 2010). For researchers with high competency (i.e., researcher's performance capacity), it is easier to publish in an international journal than for researchers with low competency. With increasing researcher's performance capacity, the probability of being able to publish in an international journal increases. This can be represented as an s-shaped exponential function, the so called *Item Characteristic Curve* (ICC), where the probability ranges from 0 (= no publication) to 1 (= publication), where 0 and 1 are only approximated and never reached (Fig 1.).

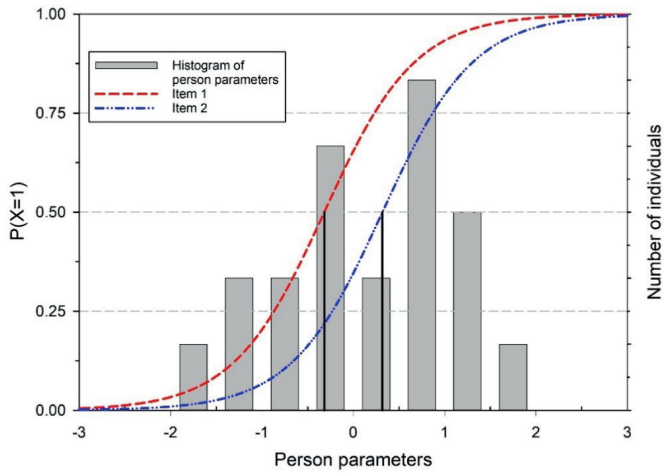


Fig. 1. Item characteristic curves for two items and histogram of person parameters (fictitious data).

However, the personal competency alone is not sufficient. The difficulty of the item must also be taken into account, as well, which represent the environment component. Thus it is much more difficult for a researcher to publish an article in a high-impact journal (e.g., “Nature”) (Item 2, Fig. 1) than in a low impact journal (Item 1, Fig. 1). Therefore, the ICC of Item 2 is shifted to the right on the x-axis in comparison to the ICC of Item 1.

In addition, items can separate differently well between researchers with high competency compared to researchers with low competency. For example, a single publication in a high-impact journal (e.g., "Nature") might already identify a researcher as excellent, since he or she then scores high in all other items as well. The item (e.g., to publish in "Nature" or not) would then have a high item discrimination as the second item parameter of the Rasch model. The s-shaped curve would be very steep (see Item 2, Fig.1).

Since bibliometric raw data (e.g., number of citations) are usually counts (i.e., integer numbers including zero), the binary Rasch model must be transformed to a count model. This is done by multiplying the binary core model with a g-component, which indicates the maximum expected count of an item in the sample. A problem is that researchers, who had more life time for their research (“active research time”) are favored over researchers with less life time, because they have more time to publish. For this reason, the active research time is also included in the model. In the last step, the model is extended to consider the impact of institutions by dividing the person parameter (histogram, Fig. 1) into two components: an institutional component and an individual-specific residual component within the institution.

Two assumptions of the model are of particular importance: Specific objectivity and local stochastic independency. According to the assumption of "specific objectivity", differences between items (e.g., item difficulties) should be independent of the sample of individuals, which were assessed, and vice versa, differences between individuals should be independent of which items are used to assess the individuals. A simple way to check this is to divide the data set into two groups (e.g., 2 subfields) and test whether the item parameters differ between the two groups. The local stochastic independency assumption assumes that the person parameters are the sole cause of the correlations among the items. If the relationships among them are statistically controlled for the person parameters, the resulting residuals are uncorrelated in the case of "local stochastic independency".

All model parameters including the person parameters can be estimated with the *two-parameter ICC Poisson counts model*, which can be formalized as follows: For $i = 1$ to N_i items as random variables X_{vi} with realized count outcomes x_{vi} for $v = 1$ to N_v researchers, the expected value in counts for the final Poisson-distributed random variable X_{vi} is (Mutz & Daniel, 2018):

$$\lambda_{vi} = E(X_{vi} \mid \boldsymbol{\varphi} = (g, \alpha_i, \xi_v, \beta_i)) = g \cdot \text{rtime}_v \cdot \left[\frac{e^{\alpha_i(\xi_v - \beta_i)}}{1 + e^{\alpha_i(\xi_v - \beta_i)}} \right] \tag{1}$$

$$X_{vi} \sim \text{Poisson}(\lambda_{vi}),$$

where g is the maximum annual value of X_{vi} (e.g., the maximal annual number of publications in the sample of individual researchers),

ξ_v is the person parameter of individual v,

β_i is the item parameter or item difficulty of item i,

α_i is the discrimination parameter for item i (the higher the value is, the more the item discriminate between individuals with high or low competency),

$rtime_v$ is the observed active research time of researcher v (the year of the last publication minus the year of the first publication of a researcher v) .

Due to the fact that the Poisson distribution is very restrictive (the mean value is equal to the variance), the Poisson distribution have often to be replaced with the Negative Binomial distribution (Mutz & Daniel, 2018; Mutz & Wolbring, 2017). In order to represent the variability between institutions, the person parameter is again divided into two components as follows (Fox, 2010, p. 145f):

$$\xi_v = \xi_{v(h)} + \gamma_h, \tag{2}$$

where $\xi_{v(h)}$ represents the individual specific component within the institution h (residual) and γ_h the effect of institution h (2-level model). The model can be estimated by a Bayesian estimation approach suggested by Stone and Zhu (2015). The within variance of $\sigma^2_{\xi v(h)}$ is fixed to 1.0 (informative prior).

Data and Methods

The publisher Elsevier provided us with bibliometric raw data from the bibliographic database Scopus to ~500,000 publications from all subject areas published between 1996 and 2015, in which at least one author with a Swiss university affiliation was involved. A comprehensive data cleansing was carried out, which mainly concerned the affiliations. The publications often use different spellings from the same institution (e.g., EPF Lausanne, EPFL, ETHL, Swiss Federal Institute Lausanne, Swiss Federal Institute of Technology Lausanne, École polytechnique féderale de Lausanne), some of which Scopus provided with a different organization ID.

Since the analysis does not primarily refer to publications, but to researchers, a sample was drawn from social scientists with the following characteristics: *experienced researchers from the social sciences, who were able to produce within 3 years at least 2 publications, which were recorded in Scopus*. According to the person ID of Scopus, researchers were selected who had published mainly according to the field classification of Scopus (ASJC) in the subfields of economics, psychology, sociology, and educational sciences (social sciences). The first publication should have been published before 2014 and publications should be available within 3 years. According to these criteria, 1,071 researchers from 12 universities and 4 subfields were selected. The combination of universities × subfields (12 × 4 = 48 and 47, respectively, since one combination was not available) was used as clusters in the multi-level model. Of the 1,071 social scientists, 291 (27.2%) were psychologists, 156 (14.6%) sociologists, 497 (46.1%) came from the economy and 127 (11.9%) from education. The academic age as the difference between the final year 2015 (time interval of the data) and the year of the first publication was on average 9.7 years (SD = 4.7) (Table 1).

The following bibliometric indicators served as items in the model: Number of top 5% publications, number of first author publications, citation of the highest cited paper, mean normalized citation score (number of papers with citation above the mean level of citations of a field), total citations for a 5-year window. A three-level statistical model was used (level 1: researcher, level 2: university × subfield, level 3: university).

Results

Sample description

The group of social scientists with Swiss university affiliations published 8.8 publications on average per capita during the study period, a minimum of 2 and a maximum of 104 (Table 2). With regard to the citation impact, 0.84 publications were on average in the top 5% percentile, about 4 publications were first author publications, about 3.4 publications were above the average of the total citations of a field. The citation of the most cited work amounted on average to 50 citations per capita. The h-index was 4 with a active research time of 7 years on average.

Table 1. Descriptive statistics per capita (N = 1,071 scientists)

Variable	Label	N	Mean	SD	Min	Mdn	P95%	Max
ITEM 1	Number of top5% publications	1,071	0.84	1.71	0	0	4	16
ITEM 2	Number of first author publications	1,071	4.05	4.50	0	3	12	42
ITEM 3	Citation of highest cited paper	1,071	49.97	108.53	1	21	169	1,639
ITEM 4	Mean normalized citation score	1,071	3.36	4.44	0	2	11	3,090
ITEM 5	Total citations (5-year window)	1,071	93.83	183.63	1	39	389	19
NPUB	Number of publications	1,071	8.82	8.94	2	6	27	104
AGE	Academic age	1,071	9.67	4.67	3	9	19	19
RTIME	Active research time	1,071	6.95	3.99	3	6	16	25
h	h-index	1,071	3.95	3.33	1	6	11	25

Note. SD = standard deviation, Min = minimum, Mdn = median, P95% = 95% percentile, Max = maximum.

With the exception of Item 2 (first authorship), psychologists had the highest mean values in all subfields. However, there were no significant differences between the subfields in the active research time.

Model comparison and model assumptions

In the first step, a model comparison was carried out to determine the model that best fitted the data (Table 2), once under the assumption of a Poisson distribution, once under the assumption of a Negative Binomial distribution.

As starting model a very restrictive one (M_1) was chosen, which assumed that all items had the same item difficulty β and item discrimination α. The restrictions were successively abandoned. The Deviance Information Criterion (DIC) served as the criterion for model comparison. The smaller the DIC, the better the model fits. In this respect, the best model was M_4, which assumes that the items have different item difficulties and discriminations. Models with a Negative Binomial distribution were clearly favored toward models with a Poisson distribution.

Of the *additional models*, a two-dimensional model (M_5) outperformed both, a model that allowed differences between subfields in the mean value of the person parameter and in the

item difficulties (M_6), and a model (M_7) that took into account the hierarchical structure of the data (3-level model). However, an additional multilevel model (M_8), in which the g-parameters were fixed in advance (informative prior), and which showed quite better convergence in the estimation process ("stationarity of Markov chains"), outperformed all other models and was selected as the final model. Eventually, a measurement model was obtained with a one-dimensional scale, where mean differences between subfields could be neglected.

Table 2. Model comparison with the Deviance Information Criterion (DIC).

	Factors				DIC	
MNo	Dimen.	Item difficulty β	Item discrimination α	Scale	Poisson	Negative Binomial
1	one	Equal	equal (=1)	equal	175,910.85	44,155.0
2	one	Unequal	equal (=1)	unequal	-	25,159.9
3	one	unequal	equal	unequal	42,132.3	25,051.6
4	one	unequal	unequal	unequal	41,952.2	**24,965.7**
Additional models						
5	two	unequal	unequal	unequal	26,124.9	24,850.4
6	M_4 + subfield differences in mean & difficulty				41,426.0	24,872.0
7	M_4 + multilevel				42,050.9	25,021.9
8	Final model: M_7 + g-components fixed				-	**24,631.4**

Note. "Equal" means that the respective item parameter value is constant cross items. "Unequal" means that the items vary in the respective item parameter. The lowest DIC values are bold faced.

Apart from the one-dimensionality, local stochastic independence is another prerequisite of the Rasch model. If the inter-correlations among the items are statistically controlled for the person parameters, the residual correlations should disappear (~0). In fact, the correlations between the items largely almost disappear, if one goes from the observed data (below diagonal) to the residuals (above diagonal) (Table 3). Thus, the assumption of local stochastic independence was widely confirmed.

The reliability of the scale amounted to r_{tt} = .84 and was rather high, but not perfect.

Table 3. Item inter-correlations (Spearman) for observed values (below diagonal) and for model residuals (above diagonal)

	Item 1	Item 2	Item 3	Item 4	Item 5
Item 1	1.00	-.01	-.02	.11	.01
Item 2	.35	1.00	-.02	.12	-.00
Item 3	.51	.17	1.00	-.02	.03
Item 4	.78	.54	.39	1.00	.03
Item 5	.79	.40	.71	.79	1.00

Model interpretation

Instead of interpreting the model parameters, two figures are chosen in order to represent the model results. As explained above, the Poisson Rasch model consists of a binary core model and a frame model applicable to count data.

In the *binary core model* (Fig. 2) the probability to score excellently and to reach the maximum value of an item (e.g., the highest possible annual citation) is related to the person parameter, i.e. the researcher's performance capacity. With increasing person parameter value, the probability to score excellently increased. The turning points of the ICC, which are linked with vertical lines (Fig. 2), indicate the item difficulties.

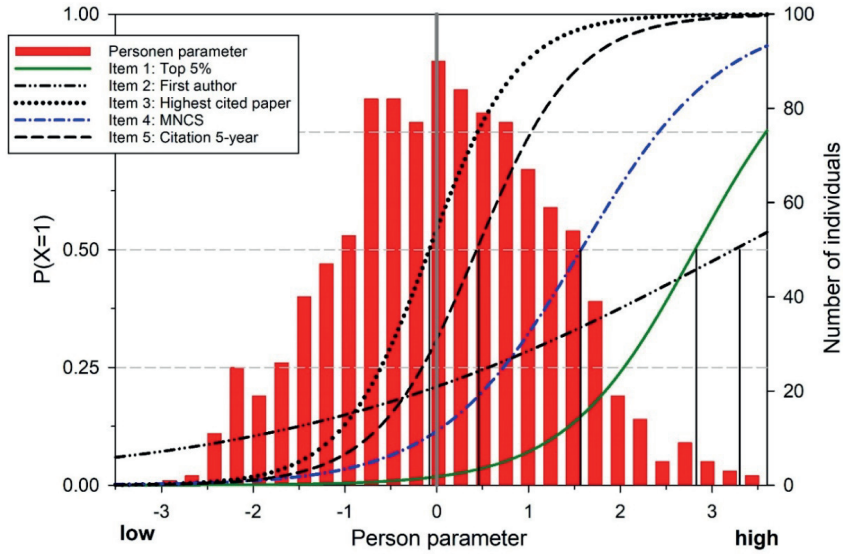


Figure 2. Item characteristic curve plot for the binary core Rasch model.

The following results can be formulated:

- *Person parameter*: Like most psychological characteristics (e.g., extroversion), the person parameters were symmetrically and normally distributed. In contrast, the bibliometric raw data are, actually, skewed distributed (e.g., Mutz & Daniel, 2012). For about 50% of the sample the person parameters were below 0 with probabilities less than 0.5 in all items. This means that half of the sample reached only half of the maximum annual rates in all items (e.g., highest citation).
- *Item parameter*: The two items for the raw citations (Item 3 and 5) showed the lowest item difficulties and the highest item discriminations of all items. It was easier for the social scientists to publish excellently in comparison to their colleagues from Switzerland (highest cited paper, citation 5-year window) than to publish excellently in international comparison regarding their field (top 5%, MNCS). The non-field normalized items distinguished better between researchers with high and low performance capacity than the field-normalized ones. Of low importance was the first authorship (Item 2), which showed both a low power to separate between researchers with high and low competency (item discrimination) and a high item difficulty. The items were more suitable for distinguishing scientists, which scored above average, than scientists, which scored below average.

The *frame model* allows the interpretation of the parameters in units of the items e.g., number of publications or citations (Fig. 3). In the ICC plot the annual accounts are related to the bibliometric quotient (BQ), which results from a simple linear transformation of the person parameters (Fig. 2) with mean value 100 and standard deviation of 15, which allows an formal not content-related interpretation of the BQ similar to an intelligence quotient (Fig. 3).

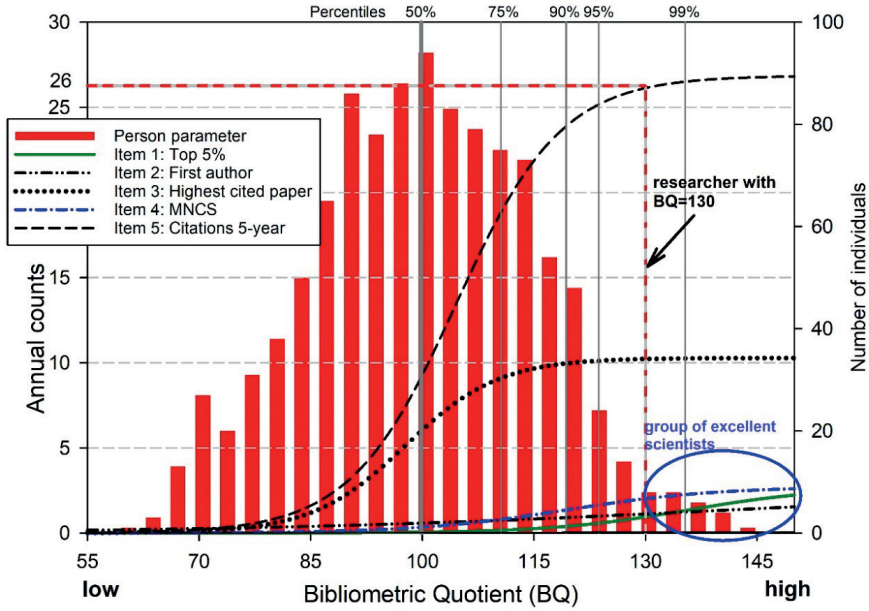


Figure 3. Item characteristic plot for the Poisson Rasch model for count data. Example: A researcher with BQ of 130 is likely to get 26 annual citations for his or her work.

The shrinkage correction of the Negative Binomial distribution and the "active research time" are not considered in the figure to facilitate the model interpretation.

According to Fig. 3 the BQ ranged from 59 (minimum) to 145 (maximum). A distinguishable group of very excellent scientists became visible, who had a BQ of over 130 (2 SD) and scored highly in all items. About 1% of the scientists had a BQ of 135 and higher.

Multilevel model

The final M_8 model also takes into account the fact that social scientists belonged to different subfields of social sciences (e.g., psychology) and different Swiss universities. The 2-level intra-class correlation (researcher, cluster) amounted to $\rho = .33$, i.e., 33% of the variability of the BQ was due to differences *between* the clusters subfields × universities (Level 2), and 67% to the variance *within* the clusters (Level 1). Only 7% of the total variance of the BQ was due to differences between universities (Level 3). The ranking of Swiss universities (Fig. 4) showed that the École Polytechnique Fédérale de Lausanne (EPFL) ranked first in the field of social sciences. However, the Goldstein-adjusted 95% credible intervals (Hox, 2010, p. 25) overlapped to such an extent that the differences in ranks could not be interpreted anymore.

Table 4. Correlations.

Correlation coefficient	Value
Intra class correlation ρ (researcher, cluster)	.33
Correlation (Spearman rank) of BQ with	
h-index	.72
total citations	.79
number of publications	.48

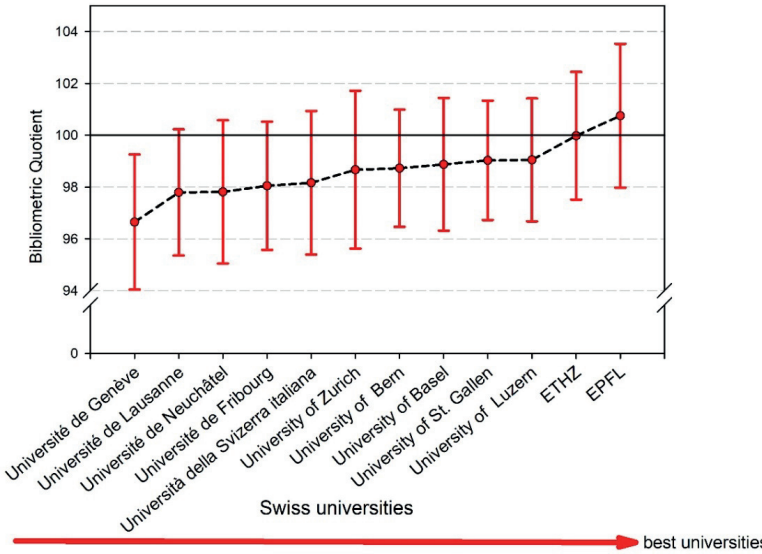


Figure 4. Ranking of Swiss universities from left to right (best universities) with Goldstein-adjusted 95% credible intervals.

Last but not least, classic bibliometric indicators such as the h-index or the total citations (Table 4) were highly correlated with the BQ (>.70).

Discussion

The amount of research articles, on how individual research performance can be measured, has increased significantly in recent years. Wildgaard et al. (2014) listed alone 108 author-level bibliometric indicators in their review. While in economy, sociology and information science the indicator approach is very common, in psychology and educational sciences scales are favored, which have to meet test theoretical requirements and have to take into account the fact that every measurement might be affected by measurement errors. It is assumed that the "researcher`s performance capacity" as trait and theoretical construct is responsible for the research output of a researcher. This can be hold for the indicator approach as well. With decisions such as the award of scholarships, it is not of primary interest which h-index a researcher has or how many top 5% articles he or she has published. Rather, it is a question of whether a researcher is able to influence his or her scientific field with his or her publications and to what extent the h-index or any other indicator or scale can say something about this competency.

The present contribution attempt to create a measurement scale on the base of test-theoretical concepts of psychology and educational sciences, which takes into account the multilevel structure of data (within and between institutions and subfields) with the following results:

- According to the Rasch model, the items formed a one-dimensional scale for assessing the "researcher's performance capacity" with a high, but not perfect reliability of r_{tt} = .84. The items were affected by measurement errors.
- Unlike bibliometric raw data the "researcher`s performance capacity" measured by the BQ was like other psychological characteristics approximately normally distributed. There was a group of very excellent social scientists with a BQ of over 130, who performed very well in all items.
- While it was easier for social scientists from Switzerland to perform well in comparison to their Swiss colleagues (raw citations), it was harder to perform well in the international arena (field-normalized citations). The items were most suitable for scientists, who scored above the average of the sample.
- Although around 33% of the variance was due to differences between the clusters subfield × university (67% within clusters), only 7% of the overall variance was actually due to differences between Swiss universities. A ranking in social sciences does not make any sense.
- The BQ is strongly related to classic bibliometric indicators, and it is not an artifact.

The results are limited, among other things, in that only a certain time interval could be used to estimate the BQ of a researcher. The results cannot necessarily be generalized to other countries. While the indicator approach at the level of institutions and countries has proved its worth, the question is whether alternative approaches at the individual level are needed that consider measurement errors. The model-oriented approach has the advantage of empirically testing certain questions of fairness, reliability, validity and invariance as empirical assumptions. A further question might be the influence of the sample selection process (only experienced researchers were focused) on the empirical results.

References

Abramo, G., Cicero, T., & D`Angelo, C. A. (2012). The dispersion of research performance within and between universities as a potential indicator of the competitive intensity in higher education systems. *Journal of Informetrics, 6*(2), 155-168. doi:https://doi.org/10.1016/j.joi.2011.11.007

Abramo, G., Cicero, T., & D`Angelo, C. A. (2013). Individual research performance: A proposal for comparing apples to oranges. *Journal of Informetrics, 7*(2), 528–539. doi:http://dx.doi.org/10.1016/j.joi.2013.01.013

Abramo, G., D`Angelo, C. A., & Grilli, L. (2015). Funnel plots for visualizing uncertainty in the research performance of institutions. *Journal of Informetrics, 9*(4), 954 -961 doi:https://doi.org/10.1016/j.joi.2015.08.006

Abramo, G., & D'Angelo, C. A. (2014). How do you define and measure research productivity? *Scientometrics, 2*(9), 1129-1144. doi:http://dx.doi.org/10.1007/s11192-014-1269-8

Alvarez, P., & Pulgarin, A. (1996a). Application of the Rasch model to measuring the impact of scientific journals. *Publishing Research Quarterly, 12*(4), 57-64. doi:https://doi.org/10.1007/BF02680575

Alvarez, P., & Pulgarin, A. (1996b). The Rasch model. Measuring information from keywords: The diabetes field. *Journal of the American Society for Information Science and Technology, 47*(6), 468-476. doi:https://doi.org/10.1002/(SICI)1097-4571(199606)47:6<468::AID-ASI7>3.0.CO;2-T

Alvarez, P., & Pulgarin, A. (1996c). The Rasch model. Measuring the impact of scientific journals: Analytical chemistry. *Journal of the American Society for Information Science and Technology, 47*(6), 458-467. doi:https://doi.org/10.1002/(SICI)1097-4571(199606)47:6<458::AID-ASI6>3.0.CO;2-U

Andrich, D. (2010). Rasch models. In P. Peterson, E. Baker, & B. McGaw (Eds.), *International Encyclopedia of Education* (3rd ed., Vol. 1, pp. 111-122). Oxford: Elsevier Science.

Bonaccorsi, A., & Cicero, T. (2016). Distributed or concentrated research excellence? Evidence from a large-scale research assessment exercise. *Journal of the Association for Information Science and Technology, 67*(12), 2976-2992. doi:https://doi.org/10.1002/asi.23539

Bornmann, L., & Marx, W. (2013). Evaluating individual researchers' performance. *European Science Editing, 39*(2), 39-40.

Bornmann, L., & Marx, W. (2014). Distributions instead of single numbers: Percentiles and beam plots for the assessment of single researchers. *Journal of the Association of Information Science and Technology, 65*(1), 206-208. doi:https://doi.org/10.1002/asi.22996

Bornmann, L., & Mutz, R. (2011). Further steps towards an ideal method of measuring citation performance: The avoidance of citation (ratio) averages in field-normalization. *Journal of Informetrics, 5*(1), 228-230. doi:https://doi.org/10.1016/j.joi.2010.10.009

Doebler, A., Doebler, P., & Holling, H. (2014). A latent ability model for count data and application to processing speed. *Applied Psychological Measurement, 38*(8), 587-598. doi:https://doi.org/10.1177/0146621614543513

Fox, J.-P. (2010). *Bayesian Item Response Modeling*. New York: Springer.

Harnad, S. (2008). Validating research performance metrics against peer rankings. *Ethics in Science and Environmental Politics, 8*(1), 103-107. doi:https://doi.org/10.3354/esep00088

Hox, J. J. (2010). *Multilevel Analysis: Techniques and Applications* (2nd ed.). New York: Routledge.

Karlsson, A., Hammarfelt, B., Steinhauer, H. J., Falkman, G., Olson, N., Nelhans, G., & Nolin, J. (2015). Modeling uncertainty in bibliometrics and information retrieval: An information fusion approach. *Scientometrics, 102*(3), 2255-2274. doi: https://doi.org/10.1007/s11192-014-1481-6

Mutz, R., & Daniel, H.-D. (2012). Skewed citation distributions and bias factors: Solutions to two core problems with the journal impact factor. *Journal of Informetrics, 6*(2), 169-176. doi:https://doi.org/10.1016/j.joi.2011.12.006

Mutz, R., & Daniel, H.-D. (2018). The bibliometric quotient (BQ), or how to measure a researcher's performance capacity: A Bayesian Poisson Rasch model. *Journal of Informetrics, 12*(4), 1282-1295. doi:https://doi.org/10.1016/j.joi.2018.10.006

Mutz, R., & Wolbring, T. (2017). The effect of the "very important paper" (VIP) designation in Angewandte Chemie International Edition on citation impact: A propensity score matching analysis. *Journal of the Association for Information Science and Technology, 68*(9), 2139-2153. doi:https://doi.org/10.1002/asi.23701

Stone, C. A., & Zhu, X. (2015). *Bayesian Analysis of Item Response Theory Models using SAS*. Cary, NC: SAS Institute Inc.

Wechsler, D. (1958). *Measurement and Appraisal of Adult intelligence* (4th ed.). Baltimore: Williams & Wilkinds.

Wildgaard, L., Schneider, J. W., & Larsen, B. (2014). A review of the characteristics of 108 author-level bibliometric indicators. *Scientometrics, 101*(1), 125-158. doi:https://doi.org/10.1007/s11192-014-1423-3

Comparing Breakthrough and Non-Breakthrough Papers from Early Citing Structures

Chao Min[1], Yi Bu[2], and Jianjun Sun[3]

[1] mc@nju.edu.cn, [3] sjj@nju.edu.cn,
Nanjing University, School of Information Management, Nanjing (China)

[2] buyi@iu.edu
Indiana University, Center for Complex Networks and Systems Research, School of Informatics, Computing, and Engineering, Bloomington, IN (U.S.A.)

Abstract

Breakthrough research plays an essential role in the advancement of science system. The identification and recognition of scientific breakthroughs is thus of extreme importance. We propose a citing-structure perspective for observing the unfolding of breakthrough research in the variations of knowledge structure. A series of network topology indictors are used to differentiate the citing networks of over 100 pairs of breakthrough papers and their control papers. 330 pairs of citing networks are subject to statistical tests for those indicators. The results show that compared with less ground-breaking papers, breakthrough papers show salient performance in terms of number of nodes, number of edges, average clustering coefficient, number of components, and average degree. Indicators such as network density, maximum betweenness centrality, and maximum closeness centrality, however, do not have a significantly discriminative power. It reveals that breakthrough papers have more connected citing networks than papers with less ground-breaking ideas. This characteristic has potential implications for early identification of scientific breakthroughs.

Introduction

Scientific breakthroughs often lead to scientific revolutions (Kuhn, 1962) that change the way we know the world. In Kuhn's time, scientific revolutions were not easily validated in empirical data. With the development of digitalization technologies, large scale scholarly data become widely available, bringing possibilities to quantitatively study scientific breakthroughs, even scientific revolutions, from the data. The aim of this study is to explore the unique features of scientific breakthroughs from the perspective of their early citing structures.

It is difficult to operationalize a precise definition of breakthrough article or breakthrough research. The definition of breakthrough is various in dictionaries, but it is considered to be associated with important discoveries and further development in many cases. For example:

- Oxford dictionary: "a sudden, dramatic, and important discovery or development."
- Collins dictionary: "a significant development or discovery."
- The Free dictionary: "a major achievement or success that permits further progress."

In addition, breakthrough is also referred to transformative research in literature. The U.S. National Institutes of Health defines transformative research as "unconventional research projects that have the potential to create or overturn fundamental paradigms"[i]. The National Science Board (NSB) also describes transformative research as resulting in "a new paradigm or field of science or engineering" (NSB, 2007).

Despite the various definitions, whether a scientific work is considered breakthrough research is eventually decided by peers (Schneider & Costas, 2017). Therefore, in this study, we select Nobel Prize winning papers as a benchmark of breakthrough research. Rather than focusing on the longitudinal depiction of individual scientific breakthroughs, we compare breakthrough papers with their non-breakthrough (or "less ground-breaking") counterparts to find their differences. We conduct such a comparison on more than one hundred pairs of such cases.

A research perspective distinct from prior studies (e.g., Ponomarev, Williams & Hackett et al., 2014; Ponomarev, Williams & Lawton et al., 2012; Schneider & Costas, 2017) is adopted here, namely a citing structure method. We define citing structure as the network topology of the citation network of the citing publications of a focus paper. In such a network, all nodes are precisely the citing publications of a paper of interest, and the citation relations among these citing publications are also considered. Huang et al. (2018) termed this network as a citing cascade. A prior study (Min, Bu, Sun, & Ding, 2018) has shown that certain features of citation patterns can differentiate breakthrough and non-breakthrough papers and the effect is most significant in the first-order citing structure. In this study, we further explore the difference in early network topology of directly citing publications between papers of different extent of breakthrough. We seek to find answers to the following research questions (RQs):

- RQ1: Does citing structure of breakthrough research show particular characteristics?
- RQ2: What particular characteristics does it show if they exist? and
- RQ3: Why does it show such characteristics if they exist?

Related Work

Observing the traces of scientific breakthroughs from bibliometric data has been pursued by researchers from many distinct perspectives (Garfield, 1977; Bettencourt, Kaiser & Kaur, 2009; Shi, Leskovec & McFarland, 2010; Shibata, Kajikawa & Matsushima, 2007; Small, Tseng & Patek, 2017; Schneider & Costas, 2017). Bettencourt et al. (2009), for example, held that the creation and spread of scientific discoveries lead to measurable changes in the social structure of a scientific community and found the evidence in the topological transitions in collaboration networks of researchers. Shi et al. (2010) looked into the citation relations among papers that a certain paper refers to, which they termed as citation projection graph. They found that how scientific papers draw previous knowledge together correlates with the paper's forward citation impacts. Shibata et al. (2007) investigated the correlation between citation counts and centrality measures like clustering centrality, closeness centrality, and betweenness centrality, among which betweenness centrality was found to be correlated with future citations.

While the attempts have been from many aspects, we in this study focus on an underexplored one, the structural properties of citing patterns. This approach is inspired by the observations that citation growth has structural dimension (e.g., Garfield, 1977, 2006; Hu & Rousseau, 2016, 2017). A significant example was pointed out by Garfield (2006) that Albert Einstein has published relatively few highly cited papers, but his work was cited by other super-cited Nobel class scientists. Another illustration of the second-generation citation effect was demonstrated in Francis Crick's work which has been cited by 50 super-cited papers.

The publication of a particular paper, therefore, not only releases a new piece of knowledge, but also changes the intellectual structure in its field of knowledge (Chen, 2012; Chen et al., 2009; Lv, Ding, Song & Duan, 2018). Leydesdorff (2001) shows that newly published scientific papers may make fundamental changes on the existing body of knowledge. This is vividly illustrated by the diagrams he devised to display how a particular paper reduces the uncertainty in the current state of knowledge. These structural changes of networks may in turn influence the spread of information over the networks (Lahiri, Maiya, Sulo, Habiba, & Wolf, 2008). In citation networks, specifically, Takeda and Kajikawa (2010) reported three stages of clustering: first the formation of core clusters, and then the emergence of peripheral clusters, finally the predominance of core clusters. Upham, Rosenkopf, and Ungar (2010) called these cohesive intellectual communities in knowledge networks as "schools of thought." Their analysis of scientific papers reveals how "schools of thought" both promote and constrain knowledge

creation. They concluded that inclusion in a school of thought is particularly advantageous for new knowledge and that the most impactful position within a school of thought is in the semiperiphery.

Based on the above intuitions, we propose a hypothesis that the more radical the breakthrough research contained in a paper, the more drastic changes it causes to the knowledge structure. We test the hypothesis by scrutinizing the citing structure of breakthrough and non-breakthrough papers. We try to find out if there does exist significant difference in the topology of the citing structure and explain the underlying mechanisms.

Methods and Data

Following a prior study (Min et al., 2018), we select papers that helped their authors won Nobel Prize as a group of breakthrough papers. Considering merely analysing breakthrough papers may not be sufficient for uncovering their peculiarity, we take one more step further by comparing them with similar papers that didn't win Nobel Prize. The citing networks of these papers are compared by their following properties and statistics.

For a network G, V are the set of nodes and E are the set of edges.
(1) Number of nodes: $|V|$, the number of nodes in a network;
(2) Number or edges: $|E|$, the number of edges in a network;
(3) Average degree:

$$\text{Average degree} = \frac{\sum_{v \in V} \deg(v)}{|V|};$$

(4) Network density:

$$Network\ density = \frac{2|E|}{|V|(|V|-1)};$$

(5) Average clustering coefficient (ACC):

$$ACC = \frac{1}{|V|} \sum_{v_i} C(v_i);$$

where $C(v_i)$ is the clustering coefficient of node v_i and is expressed as:

$$C(v_i) = \frac{number\ of\ closed\ triads\ connected\ to\ v_i}{number\ of\ triples\ of\ vertices\ centered\ on\ v_i};$$

(6) Maximum betweenness centrality: the highest betweenness centrality of nodes in G. The betweenness of a node v_i is:

$$\mathrm{B}(v_i) = \sum_{j<k} g_{jk}\,(v_i)/g_{jk}$$

where g_{jk} is the number of shortest paths from node v_j to node v_k, and $g_{jk}(v_i)$ is the number of shortest paths that pass through v_i;
(7) Max closeness: the highest closeness centrality of nodes in G. The closeness of a node v is:

$$\mathrm{C}(v) = \frac{|V|-1}{\sum_{u=1}^{|V|-1} d(v,u)},$$

where $d(v,u)$ is the shortest-path distance between nodes v and u; and
(8) Number of components: the number of connected components in a network.

As the source paper always has the highest degree in the citing network, it is excluded from the network when we calculate the indicator average degree. The same procedure is done when we calculate the number of components of the citing network.

Following Shen & Barabási (2014), we select 116 Nobel Prize winning papers (Nobel group hereafter) as breakthrough papers, and their counterparts as a control group that were published in the same journal, same year, having received approximately equivalent citation counts. The

citing network of each of these papers is extracted from an in-house version of Web of Science database. These citing networks are simplified to undirected graphs when we calculate the network metrics. As early citing structure is more useful in prediction and identification of breakthroughs, we focus on the citing networks within four years of the target paper's publication.

Results

Comparing two groups of citing structures

850 (425 pairs of) citing networks of Nobel Prize papers and their control papers are finally extracted. These networks contain information about the citing structure of the two groups of papers within one, two, three, and four years of publication, respectively. To make more reliable comparisons, citing networks of the first year are eliminated as a paper could be published between January and December in the same year. This leads to 660 (330 pairs of) networks left. The 330 pairs of network observations are subject to Wilcoxon signed-rank test for network topology indicators. Results in Table 1 answer our first research question, suggesting that there do exist significant differences in citing structure of papers with different degrees of breakthrough. We next look into the second and third research questions: within the scope of this study, in which indicators the two groups of papers differ and why.

Table 1. Wilcoxon signed-rank test results for network topology indicators

Indicator	***NoN***	***NoE***	*ND*	***ACC***	*MBN*	*MCN*	***NoC***	***AD***
Sig.	***0.043***	***0.001***	0.216	***0.000***	0.359	0.354	***0.034***	***0.000***
Nobel group Avg.	***160***	***886***	0.149	***0.463***	0.004	0.539	***23***	***4.189***
Control group Avg.	***151***	***668***	0.172	***0.384***	0.011	0.549	***30***	***2.938***

Note: NoN = Number of nodes, NoE = Number of edges, ND = Network density, ACC = Average clustering coefficient, MBN = Maximum betweenness centrality, MCN = Maximum closeness centrality, NoC = Number of components, AD = Average degree.

Distinctive citing structure indicators

In a citing network, the number of nodes (NoN) is precisely equivalent to the number of direct citations to the target paper. This indicator is only slightly different at the level of 0.05, but the average value is quite close for Nobel (160) and control group (151). This in the meanwhile indicates that the control group we select has comparable citation impacts with the breakthrough papers. Even so, Nobel group has significantly more edges (886) than control group (668) does in the citing networks (NoE). The additional edges come from more citation relationships among the citing publications of Nobel Prize papers. This is further confirmed by the finding that the citing papers of Nobel group has on average 1.2 more papers connecting to them in the citing networks than those of control group do, reflected by the indicator average degree (AD). Another significantly distinctive indicator is average clustering coefficient (ACC) of citing networks. It measures the connectedness between each of the neighbours surrounding an average node in a network. The Nobel group's citing networks show higher average clustering coefficient than that of control group, suggesting that an average node in Nobel Prize paper's citing network has more connected neighbourhood. Number of components (NoC) in citing network is also distinctive at the level of 0.05. Nobel group has visibly less components (23) than control group (30) on average in the citing networks.

Indistinctive citing structure indicators

In contrast, Nobel and control groups do not show significant difference in other citing structure indicators. These include network density (NE), maximum betweenness centrality (MBN), and

maximum closeness centrality (MCN). The effect of network density is in line with the results of a prior study (Min et al., 2018) involving higher-order citing networks, demonstrating that this indicator has a relatively weak distinctive power. We expect prominent brokerage effect in a network with high maximum betweenness, but the two groups of networks don't have significant difference in this indicator. Maximum closeness centrality, the indicator that identifies the node who has shortest distance to other nodes, is not significantly different for the two groups of citing networks, either. These indicators don't show good potential in separating Nobel Prize papers from control group.

Discussion and Conclusion

This paper reports the results of a research-in-progress study on the early citing structure of a group of breakthrough papers and their control group. Different from previous studies, we adopt a citing network perspective to explore the characteristics of breakthrough papers early after publication. Results show that there does exist significant difference in network topology of those citing networks for papers with different extent of breakthrough but similar conditions in terms of publication year, venue, and citation impact.

It reveals that breakthrough papers have more connected citing networks than papers with less ground-breaking ideas. This is reflected in various aspects of network topology. Compared with a less ground-breaking paper, the citing network of a breakthrough paper has globally more edges (NoE) while the number of nodes (NoN) has only a small difference. Within the citing network, an average node has significantly more nodes linking to it (AD), coming from works that it references and works that cite it. Both of the two sources of nodes also locate in the citing network. The neighbours surrounding an average node also cite each other (ACC) more frequently. The network topology is more cohesive (NoC), with fewer components disconnected with one another, despite more nodes in the network. All together, these findings provide empirical evidences for our hypothesis that radical breakthroughs would exhibit significant features in knowledge structure. The explanation we can provide for the observations in this study is that: the reason why a scientific work deserves to be called ground-breaking is not just that it solves a long puzzling problem or fits well with objective phenomena, but that it advances research frontiers and sets the framework for future agenda.

Since methodology papers often dominate the top of the citation count list, a citing-structure approach would have potential advantages over merely citation-ranking approach in identifying scientific breakthroughs. However, further research is needed to formulate citing structure in a normative manner and to tap more potential of this approach.

Acknowledgments

Financial support from Humanities and Social Sciences Program of the Ministry of Education (No. 19YJC870017) and the National Science Foundation of China (NSFC No. 71874077) is gratefully acknowledged.

References

Bettencourt, L. M., Kaiser, D. I., & Kaur, J. (2009). Scientific discovery and topological transitions in collaboration networks. *Journal of Informetrics*, 3(3), 210-221.

Chen, C. (2012). Predictive effects of structural variation on citation counts. *Journal of the American Society for Information Science and Technology*, 63(3), 431-449.

Chen, C., Chen, Y., Horowitz, M., Hou, H., Liu, Z., & Pellegrino, D. (2009). Towards an explanatory and computational theory of scientific discovery. *Journal of Informetrics*, 3(3), 191-209.

Garfield, E. (1977). Highly cited articles. 39. Biochemistry papers published in the 1950. *Current Contents*, 25, 5–12.

Garfield, E. (2006). Identifying Nobel class scientists and the uncertainties thereof. Retrieved June 20, 2006, from http://www.garfield.library.upenn.edu/papers/lund2006.pdf .
Hu, X., & Rousseau, R. (2016). Scientific influence is not always visible: The phenomenon of under-cited influential publications. *Journal of Informetrics*, 10(4), 1079-1091.
Hu, X., & Rousseau, R. (2017). Nobel Prize winners 2016: Igniting or sparking foundational publications? *Scientometrics*, 110(2), 1053-1063.
Huang, Y., Bu, Y., Ding, Y., & Lu, W. (2018). Number verses structure: Towards citing cascades. *Scientometrics*, 117(3), 2177-2193.
Kuhn, T. S. (1962). *The Structure of Scientific Revolutions*. University of Chicago Press.
Lahiri, M., Maiya, A., & Sulo, R. (2008). Habiba, and TYB Wolf. The Impact of Structural Changes on Predictions of Diffusion in Networks. *ICDMW'08*.
Leydesdorff, L. (2001). The challenge of scientometrics: The development, measurement, and self-organization of scientific communications. Boca Raton, FL: Universal Publishers.
Lv, Y., Ding, Y., Song, M., & Duan, Z. (2018). Topology-driven trend analysis for drug discovery. *Journal of Informetrics*, 12(3), 893-905.
Min, C., Bu, Y., Sun, J., & Ding, Y. (2018). Is scientific novelty reflected in citation patterns? *Proceedings of the 81st annual meeting of the association for information science and technology*, 55(1).
NSB (National Science Board). (2007). Enhancing support of transformative research at the national science foundation (p. 14). Arlington: National Science Foundation. https://www.nsf.gov/nsb/documents/2007/tr_report.pdf.
Ponomarev, I., Williams, D., Hackett, C., Schnell, J., & Haak, L. (2014). Predicting highly cited papers: A method for early detection of candidate breakthroughs. *Technological Forecasting and Social Change*, 81, 49–55.
Ponomarev, I. V., Williams, D. E., Lawton, B. K., Cross, D. H., Seger, Y., Schnell, J., & Haak, L. L. (2012). Breakthrough Paper Indicator: early detection and measurement of ground-breaking research. In *CRIS*.
Schneider, J. W., & Costas, R. (2017). Identifying potential "breakthrough" publications using refined citation analyses: Three related explorative approaches. *Journal of the Association for Information Science and Technology*, 68(3), 709-723.
Shen, H. W., & Barabási, A. L. (2014). Collective credit allocation in science. *Proceedings of the National Academy of Sciences of the United States of America*, 111(34), 12325-12330.
Shi, X., Leskovec, J., & McFarland, D. A. (2010, June). Citing for high impact. In *Proceedings of the 10th annual joint conference on Digital libraries* (pp. 49-58). ACM.
Shibata, N., Kajikawa, Y., & Matsushima, K. (2007). Topological analysis of citation networks to discover the future core articles. *Journal of the American Society for Information Science and Technology*, 58(6), 872-882.
Small, H., Tseng, H., & Patek, M. (2017). Discovering discoveries: Identifying biomedical discoveries using citation contexts. *Journal of Informetrics*, 11(1), 46-62.
Takeda,Y., & Kajikawa,Y. (2010). Tracking modularity in citation networks. *Scientometrics*, 83(3), 783.
Upham, S., Rosenkopf, L., & Ungar, L. (2009). Positioning knowledge: schools of thought and new knowledge creation. *Scientometrics*, 83(2), 555-581.

[i] https://factor.niehs.nih.gov/2009/october/spotlight-2010grants.cfm

Analysing technological specificities of industrial sectors using corporate patent profiles with a gravity center modelling

Pierluigi Toma[1], Massimo Frittelli[2], Antoine Schoen[3] and Patricia Laurens[4]

[1] *pierluigi.toma@unisalento.it*
University of Salento, Dept of Economic Sciences, Via per Monteroni - 73100 Lecce (Italy)

[2] *massimo.frittelli@unisalento.it*
University of Salento, Dept of Innovation Engineering, Via per Monteroni - 73100 Lecce (Italy)

[3] *a.schoen@esiee.fr,* [4] *laurensp@esiee.fr*
University of Paris-Est, ESIEE - LATTS - IFRIS, 2, bd Blaise Pascal - 93160 Noisy Le Grand (France)

Abstract

This paper investigates the possibility of developing a correspondence between the industrial sector (based on ICB classification) which is attributed to a corporation and the technological composition of this corporation's patent portfolio (based on WIPO technological fields) using a mathematical model based on gravity center.
Exploiting data characterising 1288 large corporations from the Corporate Invention Board database, we carry out a two steps analysis. In the first place we compute average patent profiles for different industrial sectors. Then, we test the discriminating power of these average patent profiles by checking to what extent the analysis of a given corporate patent portfolio makes it possible to correctly predict the industrial sector to which this corporation actually belongs.
The results show that this modelling, although providing quite precise predictive information for some industrial sectors (e.g. Healthcare, Automobiles or Chemical), does not fit for some industrial sectors which produce mainly very generic (i.e. not specific) technologies (e.g. Consumer Services or Support Services).

Introduction

Evaluating the impacts of technological and industrial policies, or identifying technological evolutions in economic sectors are central issues for innovation studies. For conducting such assessment, the economic analyses of innovation often rely on patent data to identify technologies and thus needs to link economic actors of industrial sectors and technologies. Since the 1980's, several correspondence tables have been developed for connecting patents' International Patent Classification (IPC) codes with classes of industrial sectors such as ISIC, NACE, and more (Kronz and Grevink, 1980; Everson and Putnam, 1988; Verspagen et al., 1994; Johnson, 2002; Schmoch et al., 2003; Lybbert and Zolas, 2014; Van Looy et al., 2015; Neuhäusler et al., 2017; Dorner et Harhoff, 2018). The building of these corrrespondence tables beween technology domains and industrial sectors rely on different methodologies: expertise of patent examiners, associating patent data with economic actor data through the identification of patent applicant or inventor, analysis of textual similarities between industry classification descriptions and abstracts and titles appearing in patents.

Until recently, the correspondence tables link unambiguously a given technological IPC class to a single industrial sector (Johnson, 2002; Schmoch et al., 2003; Van Looy et al., 2015). Although very useful, such tables that have been adopted by organisations such as OECD or

EUROSTAT), do not properly reflect the fine-grained complexity and diversity of the technological activities currently carried out economic actors. Recently, new correspondence tables connecting technologies across several industrial sectors propose to attribute to each industrial sector a weighted profile of technologies. Both Neuhäusler and Dorner, using different methods but relying on sets of patents with applicants or inventors from Germany, have proposed such technological profiles for NACE2 industrial sectors.

Along this avenue, we explore in this research the technology-industry link through technological profiles compiled using a large and diverse set of patents, i.e. priority patents applied worldwide by the largest and most R&D intensive industrial groups; which include the parent company and its subsidiaries. We investigate to what extent such large industrial actors, who are usually often dealing with a large spectrum of technologies, show nevertheless a specific average technological profile according to the industrial sector the parent company belong to or if the technological profiles of companies belonging to different industrial sector overlap massively.

This study is carried out in two steps. Firstly, we aim at capturing the average behaviour of each industrial sector in terms of patent profiles (or *patent mix), i.e.* its patent distribution among technological fields. To this end, we propose a gravity-center-like average patent mix indicator[1]. With such an indicator in hand, we compute the average patent mix for each industrial sector. Secondly, we aim at quantifying the discriminating impact of industrial sectors on patent profiles. Specifically, we wish to determine whether firms of different industrial sectors tend to possess significantly different patent mix compositions. To this end, we consider an Euclidean distance indicator between patent mixes. By using this distance indicator, we perform the following test: each firm is assigned to an industrial sector according to its proximity with the average patent mixes computed in the previous step. The higher the percentage of firms for which the assignment test is successful, the higher is the correlation between industrial sectors and patent mixes[2].

Data

This work is carried out using a set of 1288 worldwide firms with a patenting activity. These large firms have a sustained and intensive R&D activity and were included in the Corporate Invention Board (CIB), a patent database built in 2008 to study the inventive activities of the top corporate R&D producers. For details about the firm selection and the patent database set up, see P. Laurens et al., 2015). The firms have been allocated to industrial sectors according to the ICB classification. In order to use homogeneous industrial classes, we have designed an appropriate grouping of categories resulting in 20 industrial classes[3]. The number of firms in each class is given in Table 1. We have collected the patents applied by the firms between 2006 to 2012 (over 1076 000 patent applications). Using the WIPO classification in 35 technological

[1] An extensive review of successful use of Gravity-center-modeling can be found in the book by Eiselt & Marianov (2015).

[2] In Rahman et al. (2015) gravity centers and Euclidean distances were used for a different application, but the degree of overlapping was measured based on the distance between gravity centers, rather than on the assignment test considered in this work.

[3] Some ICB industries were split into ICB supersectors or ICB sectors. The final result distinguishes 20 different industrial classes (6 classes at the industry level, 5 classes at the supersector level and 9 classes at the sector level).

fields and fractional counting over these fields, a technological profile is assigned for each patent. Then aggregating all the patent profiles for a given firm, we calculated the overall technology profile at the firm level (patent mix).

Methodology

We consider $F \in \mathbb{N}$ firms subdivided into $S \in \mathbb{N}$ industrial sectors. Specifically, for each industrial sector $s = 1, \dots, S$, there are $F\_s \in \mathbb{N}$ firms operating in the s-th industrial sector. It follows that

$$F = \sum_{s=1}^{S} F_s$$

Each firm publishes patents in one or more of the $R \in \mathbb{N}$ existing technological fields. Let $P \in \mathbb{N}$ be the total amount of patent published by the given firms, of which $P_s \in \mathbb{N}$ are the patents published by the firms of the s-th industrial sector, for each $s = 1, \dots, S$. For $i = 1, \dots, F_s$, the i-th firm of the s-th industrial sector publishes $P_{s,i} \in \mathbb{N}$ patents, of which $P_{s,i,r} \in \mathbb{N}$ in the r-th technological field, for $r = 1, \dots, R$. It follows that

$$P = \sum_{s=1}^{S} P_s, \qquad P_s = \sum_{i=1}^{F_s} P_{s,i}, \qquad P_{s,i} = \sum_{r=1}^{R} P_{s,i,r}.$$

For the i-th firm of the s-th industrial sector we define its *patent mix* as the R-uple

$$\boldsymbol{M}_{s,i} := (P_{s,i,1}, P_{s,i,2}, \dots, P_{s,i,R}) \in \mathbb{R}^R$$

Next, we wish to define:

- a suitable *average patent mix* for each industrial sector;
- a suitable *distance* between patent mixes.

Using average patent mixes and distances between patent mixes, we wish to determine:

- whether the patent mixes of an industrial sector tend to cluster around the average patent mix of that industrial sector
- whether we can predict the industrial sector of a firm given its patent mix.

Let $s = 1, \dots, S$ be an industrial sector. We start by introducing an average patent mix for the firms of the s-th industrial sector by accounting for firm size -measured by the overall amount $P_{s,i}$ of patents published by that firm-.

First, for a firm $i = 1, \dots, F_s$, we consider the *normalized patent mix*, defined as follows

$$\boldsymbol{m}_{s,i} = (p_{s,i,1}, p_{s,i,2}, \dots, p_{s,i,R}) := \frac{1}{P_{s,i}} \left(P_{s,i,1}, P_{s,i,2}, \dots, P_{s,i,R}\right) \in \mathbb{R}^R$$

By construction, the normalized patent mix fulfils

$$0 \le p_{s,i,r} \le 1, \qquad \forall\, r = 1, \dots, R$$

$$\sum_{r=1}^{R} p_{s,i,r} = 1$$

We are now ready to introduce the *average patent mix* for the s-th industrial sector

$$\bar{\boldsymbol{m}}_s = (\bar{p}_{s,1}, \bar{p}_{s,2}, \dots, \bar{p}_{s,r}) := \frac{1}{P_s} \sum_{i=1}^{F_s} F_s \, (p_{s,i,1}, p_{s,i,2}, \dots, p_{s,i,r}) \in \mathbb{R}^R$$

By construction, the average patent mix fulfils

$$0 \leq \bar{p}_{s,r} \leq 1, \qquad \forall\, r = 1, \ldots, R$$

$$\sum_{r=1}^{R} \bar{p}_{s,r} = 1$$

The average patent mix $\bar{\boldsymbol{m}}_s$ is an indicator of the average tendential patent mix for a given industrial sector $s = 1, \ldots, S$.

With a physical analogy, the average patent mix can be interpreted as the gravity center of the firms of an industrial sector, where the spatial positions are given by the normalized patent mixes $\boldsymbol{m}_{s,i}$ and all firms have the same weight. This physical analogy suggests to use the Euclidean distance as a distance indicator between two normalized patent mixes. Specifically, given any two normalized patent mixes $\boldsymbol{m} = (p_1, p_2, \ldots, p_r)$ and $\boldsymbol{m}' = (p'_1, p'_2, \ldots, p'_r)$, their distance is given by

$$d(\boldsymbol{m}, \boldsymbol{m}') := \sqrt{\sum_{r=1}^{R} (q_r - q'_r)^2}$$

With the above framework in hand, we wish to test if the firms of a given industrial sector tend to cluster around the average patent mix of that industrial sector. Again, the physical interpretation of average patent mix is helpful. In fact, it is known that the gravity center is the point that minimizes the sum of the squared distances between itself and the given points (in this case, the normalized patent mixes of a given industrial sector), see for instance Abdi (2009). This implies that the following test holds true *for any* dataset.

$$TEST\colon \sum_{i=1}^{F_s} d(\boldsymbol{m}_{s,i}, \bar{\boldsymbol{m}}_s)^2 \leq \sum_{i=1}^{F_s} d(\boldsymbol{m}_{s,i}, \bar{\boldsymbol{m}}_{s'})^2, \qquad \forall\, s, s' = 1, \ldots, S, \qquad s \neq s'$$

However, if the firms of an industrial sector are significantly clustered around its average patent mix, the inequality in TEST is severe. In contrast, if the equality in TEST is mild, there might be many firms for which their industrial sectors cannot be clearly traced back from their patent mixes using the above model.

Now we wish, given a firm and its patent mix, to predict its industrial sector. If TEST implied

$$d(\boldsymbol{m}_{s,i}, \bar{\boldsymbol{m}}_s) \leq d(\boldsymbol{m}_{s,i}, \bar{\boldsymbol{m}}_{s'}), \qquad \forall\, s' = 1, \ldots, S$$

then we would have a deterministic criterion to determine the industrial sector of any firm, given its patent mix. Unfortunately, TEST does not imply the above inequality. However, if the given industrial sectors are sharply clustered (i.e. the inequality in TEST is severe), the above inequality is more likely to hold true. Hence, we propose the following criterion to assign firms to industrial sector. We assign a firm with normalized patent mix m to the sector $\bar{s} = 1, \ldots, S$ such that

$$CRITERION\colon\ d(\boldsymbol{m}, \bar{\boldsymbol{m}}_{\bar{s}}) \leq d(\boldsymbol{m}, \bar{\boldsymbol{m}}_{s'}), \qquad \forall\, s' = 1, \ldots, S$$

In the next section we will test this criterion with real data.

Tests and discussion

We empirically estimate the probability that $CRITERION$ correctly reconstructs the industrial sector of a firm, given its patent profile. The tests are conducted in two steps.

In a first step, we determine for each industrial sector, the empirical probability that a firm is correctly assigned to its sector. In a second step, we determine to which extent a firm of *any* sector is correctly assigned to its actual sector. The results are shown in Table 1.
As we have S = 20 industrial sectors the probability of randomly guessing the correct industrial sector of a firm is 5%. Hence, a probability of success greater than 5% indicates that the proposed test is able to extract useful information from the patent profile of a firm.
In Test 1 we consider the whole dataset (F = 1288 firms, S = 20 industrial sectors, R = 35 technological fields). $CRITERION$ was able to correctly test the 50.78 % of the given firms.
In Test 2 we consider the sub-dataset of the most inventive firms, those with at least 10 patents. This sub-dataset contains $F = 850$ firms. $CRITERION$ was able to correctly test the 54.59% of the given firms.

Table 1. Percentages of firms for which CRITERION correctly assigns the industrial sector.

Industrial sector	*Test 1*		*Test 2*	
	Firms	*Success*	*Firms*	*Success*
Oil & Gas	29	24.14 %	22	31.82 %
Health & Care	190	67.89 %	92	71.74 %
Consumer Services	43	11.63 %	17	11.76 %
Telecommunications	21	90.48 %	17	100.00 %
Utilities	31	58.06 %	18	83.33 %
Financials	22	27.27 %	6	33.33 %
Chemicals	89	77.53 %	67	85.07 %
Basic resources	36	52.78 %	31	58.06 %
Automobiles & Parts	73	68.49 %	64	76.56 %
Food & Beverage	42	69.05 %	19	78.95 %
Personal & Household Goods	82	30.49 %	60	36.67 %
Construction & Materials	39	53.85 %	29	48.28 %
Aerospace & Defense	42	38.10 %	26	42.31 %
General Industrials	43	30.23 %	40	35.00 %
Electronic & Electrical Equipment	127	51.97 %	96	33.33 %
Industrial Engineering	114	33.33 %	77	41.56 %
Industrial Transportation	8	25.00 %	2	50.00 %
Support Services	21	9.52 %	9	22.22 %
Software & Computer Services	83	60.24 %	27	75.68 %
Technology Hardware & Equipment	153	45.75 %	121	49.59 %
Overall	1288	50.78 %	850	54.59 %

The average patent mixes of the considered dataset reveals that all the industrial sectors possess a non-negligible amount of patents in every technological field. In other words, there are no technological fields that uniquely characterize a specific industrial sector. For this reason, the authors believe that developing more viable tests for the automated detection of industrial sectors, based only on the knowledge of patent profiles, is bound to limited success. However, the results also show huge variations in the empirical probability depending on the

industrial sectors. In a few sectors (Healthcare, Automobiles or Chemical), the probability to properly assign the industrial sector of a firm given its patent technological profile is high (above 70%). In such cases, this methodology could be used to predict either the industrial sector of a given patent profile of a firm or the patent technological portfolio of a firm knowing its industrial sector.

Future directions

In a next step, we will study if we can improve our results by considering not only the one and only industrial sector which the firm belongs to but by analyzing as well the more complete firm's industrial profile that reflects the distribution of the firm's subsidiaries over industrial sectors. We will test the possibility to calculate a theoretical technological patent profile for a firm taking into account its full industrial profile. If successful, such a methodology could be very useful when dealing with data quality when updating corporate patent database.

Moreover, since the proposed gravity-center-like analysis shares some similarities with the problems of clustering and weighted clustering, (see for instance the work by Ackerman et al., 2012), using suitable clustering techniques could be also a future development of this work.

Acknowledgments

This work was supported by RISIS, a project of the Infrastructure EU FP7 programme.

References

Ackerman, M., Ben-David, S., Brânzei, S. & Loker, D. (2012). Weighted Clustering. In *Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence* (pp. 858-863). AAAI Press.

Abdi, H. (2009). Centroids. *Wiley Interdisciplinary Reviews: Computational Statistics*, *1*(2), 259-260.

Eiselt, H.A. & Marianov, V. (Eds.), *Applications of Location Analysis*. Springer.

Everson, R., Putnam, J., 1988. The Yale-Canada patent flow concordance. *Economic Growth Centre Working Paper*. Yale University, New Haven.

Johnson, D.K.N., 2002. The OECD Technology Concordance (OTC). Patents by Industry of Manufacture and Sector of Use. *OECD STI Working Papers 2002/5*. OECD, Paris.

Kronz, H., Grevink, H., 1980. Patent statistics as indicators of technological and commercial trends in the member States of the European Communities (EEC). *World Patent Inf.* 2 (1), 4–12.

Laurens, P., Le Bas, C., Schoen, A., Villard, L., Larédo, P., *Research Policy,* 44, (2015) 1763–1772.

Lybbert, T.J., Zolas, N.J., 2014. Getting patents and economic data to speak to each other: an 'Algorithmic Links with Probabilities' approach for joint analyses of patenting and economic activity. *Res. Policy* 43 (3), 530–542.

Neuhäusler, P., Frietsch, R., Mund, C., Eckl, V., 2017. Identifying the technology profiles of R&D performing firms—a matching of R&D and patent data. *Int. J. Innov. Technol. Manage.* 14 (1).

Rahman, A.I.R.J., Guns, R., Rousseau, R., Engels, T.C.E., 2015. Is the expertise of evaluation panels congruent with the research interests of the research groups: A quantitative approach based on barycenters. *Journal of Informetrics* 9 (4), 704-721.

Schmoch, U., Laville, F., Patent, P., Frietsch, R., 2003. Linking Technology Areas to Industrial Sectors. *Final Report to the European Commission, DG Research.* Karlsruhe, Paris, Brighton.

Van Looy, B., Vereyen, C., Schmoch, U., 2015. Patent Statistics: Concordance IPC V8–NACE Rev.2 (version 2.0). *EUROSTAT*, Luxembourg.

Verspagen, B., van Moergastel, T., Slabbers, M., 1994. MERIT Concordance Table: IPC –ISIC (Rev.2). MERIT Research Memorandum 2/94-0. Maastricht Economic Research Institute on Innovation and Technology, Maastricht.

Are Special Issues that Special?
Distinctiveness and Impact of Special Issues in LIS Journals

Maxime Sainte-Marie[1], Philippe Mongeon [2] et Vincent Larivière[3]

[1] *sainte-marie.maxime@uqam.ca*
Centre Interuniversitaire de recherche sur la science et la technologie, Université du Québec à Montréal
Canada research chair on the transformations of scholarly communication, Université de Montréal

[2] *philippe.mongeon@gmail.com*
Danish Center for Studies in Research and Resarch Policy, Aarhus University

[3] *vincent.lariviere@umontreal.ca*
Canada research chair on the transformations of scholarly communication, Université de Montréal
Observatoire des sciences et des technologies, Université du Québec à Montréal

Abstract
Whether as periodical conference spinoffs or extraordinary journal digressions, Special Issues, are a common yet unexplored area of scholarly communication. In this research, content and citation analysis of Special Issues in Library & Information Science Journals indexed in the Web of Science shows that special issues are distinct from regular issues "contentwise", yet indiscernible "citationwise", thus leaving their existence and persistence unexplained from a publisher's economic perspective.

Introduction
As scholarly communication is in most fields based on journal publications, Special Issues (SIs) play a significant and lasting role in both knowledge production and dissemination. A SI can be defined as a journal issue "either completely or partly devoted" to a single topic" (Olk & Griffith, 2004, p. 120), the latter either referring to an area of study, a theoretical approach or a methodology (Priem, 2006). Despite their ubiquity, SIs do not make unanimity within the scholarly community, especially as regards to whether or not their publication is detrimental to research impact (Conlon et al., 2006; Hendry & Peichel, 2016; McKinley, 2007; Mowday, 2006; Olk & Griffith, 2004; Schoonhoven, 2004; Siguffi, 2011). At first glance, both perspectives seem plausible. On the one hand, as SIs grant "increased legitimacy and attention" (Conlon et al., 2006, p 859) to relevant or unusual topics of interest, which helps extend the journal readership and potentially boost its citation rates. Inversely, in order "to either meet deadlines or to just fill budgeted pages", journal editors may be forced to accept substandard papers, thus reducing the total number of citations received and "damaging the image of the journal" (Siguffi, 2011, p. 306).

Some empirical studies were conducted on the matter. Based on an analysis of journal issues published between 1988 and 1999 in 5 management journals, Olk and Griffith (2004) show that SI articles have a higher citation rate than regular issue (RI) articles. Expanding on this study, Conlon et al. (2006) show however that this citation boost is only apparent in lower-impact journals. Outside the field of management, Hendry and Peichel (2016) collected citation data of articles published in conference-based SIs published by the International conference on Stickleback Behaviour and Evolution. Their analysis shows that papers published in SIs have comparable citation impact and longevity to articles published in the same journal and year, as well as a lower but longer citation impact than topic-related papers published in RIs the same year. More recently, Sala, Lluch, Gil, and Ortega (2017) analyzed 1120 articles published in 10 Ibero-American psychology journals included in the Journal Citation Index and published between 2013 and 2015. By comparing RI articles to ones published in "open call" or invitation-

based SIs, the authors observe that SI papers receive a higher number of citations than RI articles, and that this higher citation impact is not the consequence of author or journal self-citations (Sala et al., 2017).

While these studies mostly agree on the research impact of SIs in their respective disciplines, they unfortunately suffer from two common shortcomings. First, their results are based on a rather small sample of issues. But most importantly, they take for granted what might be the most obvious and characteristic feature of SIs: their topicality. Regardless of discipline and whether based on open calls, conference presentations or invitations to publish, all SIs focus by definition on a more or less specific theme. And as with research impact, this topicality of SIs is not only far from trivial, but also and still in need of a proper bibliometric assessment. In light of these considerations, the aim of the present paper is to attempt a large-scale investigation of the topicality and impact of SIs. In the first step, vector semantic models are generated in order to assess the topicality of both SIs and RIs and thus verify whether or not SIs stand out in this respect. Following this, a citation analysis similar to but broader than the above-mentioned one is undertaken in order to determine whether publishing special issues contributes to a scholarly journal's influence and outreach.

Methodology

We retrieved from the Web of Science (WoS) all articles, notes and reviews published in the last 10 years (between 2009 and 2018 inclusively) in Library & Information Science journals, as classified by the National Science Foundation. We chose 2009 as starting year for our study because SIs in LIS journals are very scarce in WoS before that date. We further limited our dataset to journals that have published at least one special issue and that published at least eight issues with at least four articles each over the whole period studied. All articles that are not written in English were also removed. These dataset restrictions were applied in order to allow for reliable and robust similarity computations and comparisons. The resulting dataset contains a total of 14 132 documents published in 1 335 issues distributed amongst 34 distinct journals; of the lot, 122 (9.14%) issues and 1 213 (6.96%) articles are of the special kind. For each relevant article entry, the following attributes were extracted: article ID, title, journal, publication year, issue ID, special issue status as well as total citations, normalized by year and discipline.

Issue-level data was then obtained through the following processing tasks. Field-and-year-normalized relative citations were first aggregated at the issue level, then divided by the number of articles contained in each issue. Also, in order to obtain content similarity scores for both RI and SI articles, article title and abstract data for all collected LIS articles were merged into one single text string attribute, then segmented in vectors of 3-grams (substrings of 3 characters) with TF-IDF-weighted dimensional values. The main reason for using word substrings instead of whole words is that it allows semantically-related words such as 'science', 'scientific', 'scientifically' and 'scientist' to have non-zero similarity scores. This character sequence segmentation procedure has also been shown to offer comparable results to traditional word-based approaches over various Natural Language Processing-based tasks (Cavnar & Trenkle, 1994; Damashek, 1995; McNamee & Mayfield, 2004). Following these data processing steps, the topicality or thematic cohesiveness of each issue was obtained by calculating the average cosine similarity between the text vectors of all articles included in the issue.

Analysis

Table 1 shows the distribution by year of all LIS RIs and SIs. In the case of RIs, if we exclude journal issues published in 2018 (year for which data has yet to be collected), the number of different LIS journals having published special issues over the year remains relatively stable, while the number of issues and articles slightly increases over the period. More surprising however is the case of SIs: while the number of journals collecting special issues has remained relatively constant over time, there has been an impressive surge in number and proportion of SIs. Indeed, while counting for around 3% of all publications in 2009, the share of SIs and special articles increased to around 14% and 13% of their respective publication types in less than 10 years.

Table 1. Descriptive Stats for LIS Journals, 1991-2018.

		Special Issues		Articles		Regular Issues	Articles
Year	Journals	#	%Total	#	%Total	#	#
2009	31	4	3,2	25	2,4	120	1014
2010	31	8	5,8	65	5,5	130	1113
2011	32	12	8,3	112	8,9	133	1153
2012	31	13	8,9	99	6,4	133	1443
2013	32	15	9,9	93	7,1	137	1214
2014	33	15	8,6	132	7,7	159	1589
2015	32	15	9,0	110	6,6	152	1555
2016	31	14	7,9	112	5,8	163	1831
2017	30	17	9,9	154	8,3	154	1699
2018	25	9	14,3	81	13,1	54	538

A number of reasons can be put forward to explain this recent surge in the publication of SIs. The simplest one is that SIs were not indexed as such in the WoS until then. If that were the case, however, the relative importance of SIs would probably have increased suddenly rather than over several years. Another possible explanation may be that journal editors tend more and more to find and appoint special editors to specific issues in order to reduce their own workload. One could also suggest that more and more organization committees try to attract submissions to conferences by announcing that some or all papers chosen by the program committee will be published in a SI of a given journal. Beyond these speculations, one thing is certain though: this surge in publication and relative importance of SIs certainly warrants further investigation, be it exploratory and cross-disciplinary or explanatory. More on that matter will be said in the Discussion section.

As regards to content similarity, the percentile rank distribution of content similarity scores for both RIs and SIs is presented in Figure 1. Both distributions are positively skewed, with a little more than 18% of the highest-scoring SIs and RIs accounting for slightly more than half of the total cumulative score of their respective issue types. Additionally, while both distributions have short heads and tails, similarity scores for SIs tend to be generally higher than those of RIs. Comparison of mean content similarity scores for both types of issues further confirms this intuition, as articles within a SI are on average 17% more similar in content (0.33 vs. 0.28) than articles from any given RI. Comparison of median values leads to the same conclusion, as the median for the SI distribution (0.32) is 18% higher than that of its counterpart (0.27). Given the similar shapes of both distributions and in order to confirm their distinctiveness as regards to content similarity, we conducted a Mann-Whitney U test (Mann and Whitney, 1947) on the

similarity distributions of SIs and RIs and obtained a U-statistic of 37,970 and a p-value of 3.51×10^{-19}. Since the null hypothesis of that test states that the random variables corresponding to the two independent groups being compared are stochastically equal ("each datum of the first group will have an equal chance of being larger or smaller than each datum of the second group" (Nachar, 2008)), rejecting this hypothesis due to the very low p-value obtained entails that elements of both groups belong to distinct populations (Nachar, 2008). In the present one-tailed context, this discrepancy in central tendencies means that the elements as well as the median of one group, namely the SIs similarity score distribution, are significantly higher than those of the other group. Point biserial correlation (Lev, 1949) further confirms the higher content similarity of SIs over RIs, with r and p values of 0.26 and 5.74×10^{-22} respectively. In sum, SIs in Library and Information Sciences, whether conference-related or not, are not only a recent emerging phenomenon, but also one that also has to be distinguished from RIs "contentwise", as the articles they contain tend to form a semantically more consistent whole.

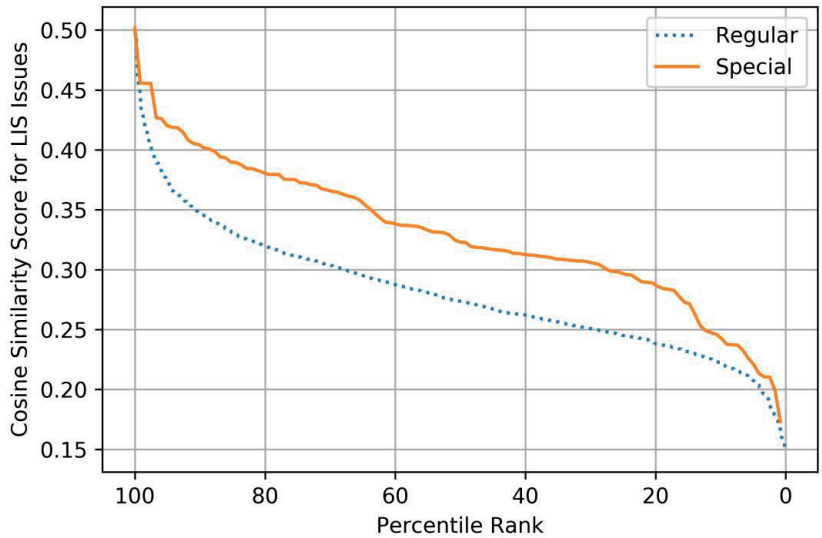


Figure 1. Frequency-Rank Distribution of Cosine Similarity Scores for LIS Issues

As regards to research impact, the frequency-rank semi-logarithmic distributions for both special and regular LIS issues are shown in Figure 3. At first glance, both distributions have strikingly similar shapes, slopes, heads, and tails. As can be expected from previous bibliometric literature, the two distributions are highly skewed, with 2.85% and 4.3% of regular and special issues accounting for more than half of the relative citations of their respective issue types. Both distributions have strikingly similar shapes, slopes and heads. As regards to the central tendencies observed for both distributions, the mean relative citation score by article value for SIs (1.18) is only 4% higher than that of the RIs distribution (1.13), whereas the median value for RIs (0.88) is 5% higher than the value for SIs (0.83). Resorting once more to point biserial correlation, we obtained a negligible r value of 0.01 for the relationship between special issue status and relative citation score per article; however, a high p-value of 0.67 prevents us from drawing any statistically significant conclusion on that matter. The Mann-

Whitney test however provides more adamant results, with U- and p- values for both RI and SI citation score distributions of 70422 and 0.19 respectively. Based on the high p-value, the null hypothesis cannot be rejected, which means that the random variables corresponding to each distribution are stochastically equal, and thus that the elements and the medians of both groups are statistically undistinguishable. These results thus suggest that the citation cost of publishing articles in SIs rather than in RIs tends to be negligible.

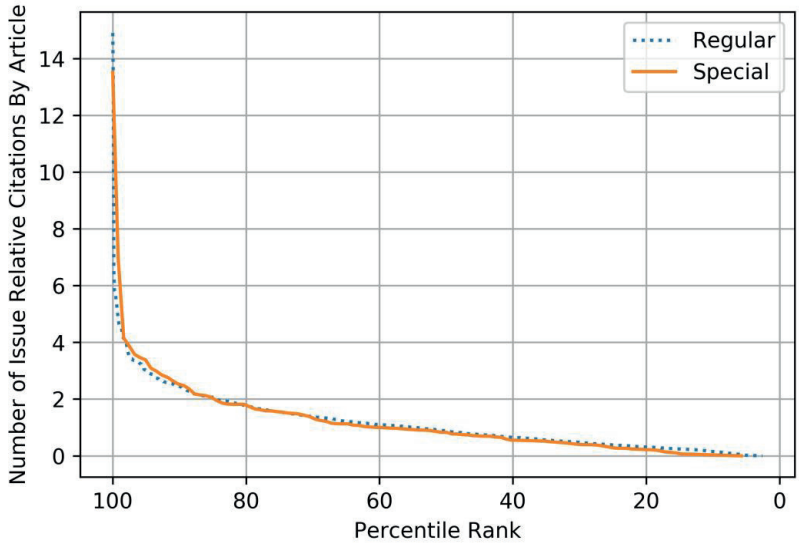


Figure 2. Frequency-Rank Distribution of Relative Citations by Article for LIS Journal Issues

Discussion

As the previous section has shown, RIs and SIs in LIS journals are distinct "contentwise", yet undiscernible "citationwise". Taken together, conference-related and unrelated SIs thus distinguish themselves from RIs from a thematic standpoint, but this specificity does not impede nor boost their research impact. Even though the explicit topicality of SIs may attract new readers or discourage others, these readership dynamics result in a zero-sum citation game. And while it is not unreasonable to suppose that some special editors may have, at times and for thematic cohesiveness purposes, accepted lesser-quality articles in order to complete their issue, nothing in the data suggests that this practice may have had an effect on other articles from the same issue or special issues in general.

Given this impact neutrality of SIs, why publish them at all then? And why are they increasing in number? In our opinion, there are necessarily incentives to publish SIs, but these benefits have to lie beyond the quantitative realm of bibliometrics. It was earlier hypothesized that journal editors may appoint special editors in order to reduce their workload. SIs may even be proposed in order to attract submissions to conferences, as mentioned before. Editing special issues or publishing in them may help gain recognition and strengthen bonds within a more local community of peers and collaborators. Editing a special issue can also be more enjoyable than editing a regular one: for the editor, "the collection of papers in a special issue can be more interesting; the review process is more collegial, constructive, and efficient; editorial decisions

are more enjoyable; and the opportunity to advance the field is greater" (Hendry & Peichel, 2016: 141). Finally, SIs are arguably more enjoyable to read than regular ones, as all articles contained within a given SI are potentially of higher relevance to the interested reader. However, given that most articles can now be directly accessed electronically, independently of any consideration at the issue level, one may arguably wonder whether speaking of issue readability or attractiveness is still relevant at all. And while all these conjectures are plausible, no scenario that would directly benefit journals and compensate for the opportunity cost mentioned in the introduction can be reasonably thought of, thus leaving the sudden and surprising increase of SIs in LIS journals without a proper and rational economic explanation.

Of course, various unaccounted factors might have affected the results presented here. For once, the quality of the database used in this research might also be questioned on various grounds. First, the accuracy of the classification of LIS issues as RIs or SIs in the Web of Science cannot be exhaustively assessed. As regards to the text-based methods used here, similarity computations between articles are entirely dependent on the wording behavior of authors, which is inherently subjective and as such often elusive to scientific inquiry. However, these limitations do not invalidate the results presented: SIs in LIS represent a scholarly communication form that is both distinct "contentwise" and indiscernible "citationwise"; whether this situation also applies to other disciplines is a matter for future research.

References

Cavnar, W.B. & Trenkle, J.M. (1994). N-gram-based text categorization. In Proceedings of SDAIR-94, 3rd Annual Symposium on Document Analysis and Information Retrieval (Las Vegas, NV, 1994), 161-175.

Conlon, D. E., Morgeson, F. P., McNamara, G., Wiseman, R. M., & Skilton, P. F. (2006). From the editors: Examining the impact and role of special issue and regular journal articles in the field of management. Academy of Management Journal, 49(5), 857-872.

Damashek, M. (1995). Gauging Similarity with *n*-Grams: Language-Independent Categorization of Text. Science, 267: 843-846.

Hendry, A.P. & Peichel, C.L. (2016). Are conference special issues worthwhile? Evolutionary Ecology Research 17: 141-155.

Lev, J. (1949). The point biserial coefficient of correlation, Annals of Mathematical Statistics 20(1): 125-126.

Mann, H.B. & Whitney, D.R. (1947). On a test of whether one of 2 random variables is stochastically larger than the other. Annals of Mathematical Statistics 18: 50-60.

McKinley, W. (2007). Special issues as vertical integration: A rejoinder to priem and mowday. Journal of Management Inquiry, 16(3), 240-245

McNamee, P. & Mayfield, J. (2004). Character *N*-Gram Tokenization for European Language Text Retrieval. Information Retrieval 7, 73-97.

Mowday, R. T. (2006). If special issues of journals are not so special, why has their use proliferated? Journal of Management Inquiry, 15(4), 389-393

Nachar, N. (2008). The Mann-Whitney U: A Test for Assessing Whether Two Independent Samples Come from the Same Distribution. Tutorials in Quantitative Methods for Psychology, 4, 1: 13-20.

Olk, P. & Griffith, T. L. (2004). Creating and disseminating knowledge among organizational scholars: The role of special issues. Organization Science, 15(1), 120-129.

Priem, R. L. (2006). What happens when special issues just aren't "special" anymore? Journal of Management Inquiry, 16(3), 246-249.

Schoonhoven, C. B. (2004). Putting rumors to rest: the impact, quality, and speed of journal special issues. Organization Science, 15(1), 119-119

Schütze, H. (1993). Word space. In S. J. Hanson, J. D. Cowan, & C. L. Giles (Eds.), Advances in neural information processing systems (Vol. 5, pp. 895-902). San Mateo, CA: Morgan Kauffmann.

Siguffi, S. P. (2011). On special issues. Journal of African Business, 12(3), 305-307.

The communication value of English-language academic journals published in non-native English countries: from a perspective of citation analysis

Zhenglu Yu[1] and Zheng Ma[2] and Haiyan Wang[3]

[1] luluyu@istic.ac.cn
Institute of Scientific and Technical Information of China, Beijing (China)

[2]mazheng@istic.ac.cn
Institute of Scientific and Technical Information of China, Beijing (China)

[3]wanghy@istic.ac.cn
Institute of Scientific and Technical Information of China, Beijing (China)

Abstract

English-language academic journals play an important role in scientific research and communication around the world. Especially in non-native English countries, there is strong demand to communicate with outside in academic fields. To measure the communication value of English-language academic journals published in non-native English countries, we apply the method from linguistic studies which is used to measure the communication value of languages. An indicator Q-value is designed to evaluate the English-language academic journals, which is comprised of three independent indicators: prevalence index, centrality index and international diffusivity index, which form a three-dimensional coordinate system and each journal is defined as a vector and the Q-value is the length of the vector. A total of 84 English-language academic journals are studied and the result shows that the Q-value can correctly reflect the communication value of these journals.

Introduction

We all believe that science is no border. As the main vehicle and bridge of science communication, academic journals play an important role in science exchanging. According to the theory of science centre transfer, the science centre has changed from Germany to the United States (Mintomo, 1979; Zhao, 1984; Feng 2000). The main communication language is different with the transfer of science centre which affects the language choosing of academic journals. English is the common language in the field of science and technology around the world. Most academic journals, especially the journals with high impact are published in English (Wang, 2007).

Science and Technology collaboration is an important model. With the development of S&T and globalization, scientists are more likely to join in the collaboration especially international cooperation and the co-authored papers are keep growing (Beaver,1978; Narvaez, 1995; Pao, 1982; Yu, 2017). Science publishing and communication has changed from National Science Model to Transnational model

(Zitt, 1998). The predominance of English as the global language for communication is nowadays unquestioned.

Language bias really exists (Crocker, 2010; Van, 2011). Some studies show the number of English articles is increasing in a lot of fields and this trend also been seen in the number changing of academic journals (Daniel,2006;). Authors are more likely to publish their studies in English-language journals (Egger, 1997) and show favourable attitude towards the use of English for academic publication even though their native language is not English. Publishing in other than English always means restricting knowledge dissemination (Burgess，2002).

Articles and journals in English are more accessible to larger communities and reaching more audience (Bennett, 2015) that will turn into higher citations and rankings. The serious language bias has significant influence on citation-based rankings (Van, 2011), such as impact factor rankings (Gonza´lez-Alcaide, 2012). The non-English-language journals or journals with a high proportion of non-English-language articles suffer from low citations and impact factor (Liang, 2013; Liu, 2018; Mongeon, 2016). More and more publishers tend to found international journals and make their journals and articles visible to a wider audience, gain more recognition and achieve international status. As we know, few non-English-language journals are included in Journal Citation Report and Scopus and even that are included still has lower impact factors compared with English-language journals.

Ulrichsweb include more than 300,000 periodicals (also called serials) which in all kinds of types around the world. There are 82,647 active academic journals whose key feature is “peer-reviewed” in Ulrichsweb and the total number of English journals is 64,830 which accounts for 78.4%. English is the lingua franca in the field of academic publishing around the world. More and more non-native English countries are aware of the importance of English journals which are more easily to achieve international status (Huh, 2014; Solovova, 2018; Moskaleva, 2018). In 2018, there were 373 English journals (registered in China) published in China, which was up 21.5% from the previous two years (ISTIC, 2018).

Methodology

This study borrows the design ideas of language communication value from Abram de Swaan. We define Q-value as the new index to evaluate the English-language journals published in non-native English countries. Swaan uses the concept of super central languages, central languages and peripheral languages to describe the relations of languages in the world (Abram, 2008). In this study, the Q-value is comprised of three independent indicators which are prevalence index, centrality index and international diffusivity index. Q-value forms a three-dimensional coordinate system and each journal is defined as a vector and the Q-value is the length of the vector. Through quantitative method, this study will show the communication value of English-language journals published in non-native English countries and their status in certain category.

The Q-value of English-language journals published in non-native English countries

is defined as:

$$Q_{ij} = \sqrt{P_i^2 + C_{ij}^{\ 2} + D_j^{\ 2}} \qquad (1)$$

Where Q_{IJ} is the Q-value of journal j in category i. P_i is the English prevalence of category i, and C_{ij} is the academic impact centrality of journal j in category i, and D_J is the international diffusivity of journal i's academic impact.

Discipline-English-prevalence P. People use one language more, the communication value of it will be more higher that is the prominent feature of language. In certain research discipline, English-language journals are more important when English is more prevalent or used widely. The English prevalence degree in certain discipline can be measured by the degree of English references being used. The quantity and ratio of English references to total references can represent the demand of English information. To subject i, the prevalence index is defined as:

$$P_{\mathrm{i}} = \frac{N_{i,E}}{N_{i,tot}} \qquad (2)$$

Where P_i is the prevalence of discipline i. $N_{i,E}$ is the total number of English references in discipline i. $N_{i,tot}$ is the total number of references in discipline i.

Journal's academic impact centrality in certain category C. Journals' academic impact centrality present its authority and the status in discipline. The mutual citation network in certain discipline can show journals position in subject and the communication status of one certain journal and others (Ma, 2012). The ratio which is citations by other journals in certain subject to the maximum value of all journals in this discipline can present certain journal's authority and centrality.

$$C_{\mathrm{ij}} = \frac{N_{\mathrm{ij,\ o}}}{\max_{\mathrm{j=1\sim n_i}} N_{\mathrm{ij,\ o}}} \qquad (3)$$

Where C_{ij} is the academic impact centrality of journal j in discipline i. n_i is the number of journals in subject i. $N_{ij,o}$ is the citations by other journals in subject i of journal j. In this study, in order to reflect the communication value of one journal, we didn't take self citation into account although the phenomenon of proper self citations is reasonable.

Journal's international diffusivity of academic impact D. The communication value of English-language journals includes not only the value from domestic database but also include that from international. International diffusivity can be shown through the relation of citations from domain and international.

$$D_J = \frac{N_{\mathrm{j,\ I}}}{N_{\mathrm{j,\ I}} + N_{\mathrm{j,\ D}}} \qquad (4)$$

Where D_j is the international diffusivity of journal i's academic impact. $N_{j,I}$ is the citations form international database. N_{jD} is the citations form domestic database.

In this study, we use English-language journals published in China as our samples. 84

journals are indexed which are included in both JCR2017 (Journal Citation Report) and CSTPCD2017 (Chinese S&T Papers Citation Database) . We use the data from these two databases to measure the Q-value of these journals. The discipline of journals are based on the criteria of reference (ISTIC, 2018) and the retrieve year is "2017".

Results

The Q-value of 84 English-language journals published in China was shown as table 1. The maximum Q-value is 1.335, the minimum Q-value is 0.459, and the mid-value is 1.173, the average Q-value is 1.170. Compared with 2012, the whole Q-value level has increased (Ma, 2012). From the distribution of Q-value, the results of table 1 show that 10 journals' Q-value above 1.3000, account for 11.9%; 23 journals' Q-value between 1.200 and 1.300, account for 27.4%; 35 journals' Q-value between 1.100 and 1.200, account for 41.7%; 14 journals' Q-value between 1.000 and 1.100, account for 16.7%; 2 journal's below 1.000, account for 2.4%. The distribution shape of Q-value is close to normal distribution.

84 journals come from the fields of fundamental research, life sciences and engineering research. The Q-value difference is not remarkable in discipline.

Some journals with distinctive characteristics have high academic influence whose Q-value shows good result, such as *Light Science & Applications, Genomics Proteomics & Bioinformatics, Cell Research, Nano Research.* Meanwhile, some journals with distinctive regional characteristics or belongs to a relatively close discipline shows a low Q-value, such as *Petroleum Exploration and Development, Chinese Medical Journal and Pedosphere.*

The Q-value of English-language journals published in China can differentiate the journals communication value and effect in Chinese S&T research activities and publishing activities.

Table 1. T-value of 84 English-language journals published in China (data 2017)

Journal	*Discipline*	Q_{ij}	P_i	D_j	C_{ij}
Acta Biochimica Et Biophysica Sinica	Bilology basic subject	1.311	0.883	0.819	0.159
Acta Mathematica Scientia	Mathematics	1.169	0.774	0.789	0.558
Acta Mathematica Sinica English Series	Mathematics	1.141	0.774	0.838	0.994
Acta Mathematicae Applicatae Sinica	Mathematics	1.227	0.774	0.836	0.276
Acta Mechanica Sinica	Mechanics	1.163	0.711	0.807	0.385
Acta Metallurgica Sinica	Metal Materials	1.175	0.708	0.669	0.159
Acta Oceanologica Sinica	Oceanography & hydrology	1.152	0.675	0.773	0.264
Acta Pharmaceutica Sinica B	Pharmacy	1.140	0.558	0.857	0.039
Acta Pharmacologica	Pharmacy	1.131	0.558	0.848	0.100

Sinica					
Advances in Atmospheric Sciences	Atmospheric Sciences	1.007	0.436	0.718	0.290
Asian Journal of Andrology	Sexual Medicine	1.107	0.607	0.829	0.414
Cancer Biology & Medicine	Oncology	1.279	0.806	0.887	0.058
Cell Research	Bilology basic subject	1.252	0.883	0.888	1.000
Chemical Research in Chinese Universities	Chemistry	1.232	0.862	0.686	0.339
China Communications	Communications Technology	1.139	0.677	0.737	0.288
China Foundry	Machine Building Equipment and Technology	1.043	0.376	0.648	0.063
China Ocean Engineering	Oceanography & hydrology	1.174	0.675	0.589	0.076
Chinese Annals of Mathematics Series B	Mathematics	1.205	0.774	0.832	0.429
Chinese Chemical Letters	Chemistry	1.275	0.862	0.810	0.296
Chinese Geographical Science	Geography	1.097	0.490	0.657	0.045
Chinese Journal of Aeronautics	Aeronautica & Aerospace	1.038	0.514	0.679	0.268
Chinese Journal of Cancer	Oncology	1.173	0.806	0.668	0.397
Chinese Journal of Cancer Research	Oncology	1.184	0.806	0.610	0.288
Chinese Journal of Catalysis	Chemistry	1.232	0.862	0.683	0.335
Chinese Journal of Chemical Engineering	Chemical Engineering,Multidicipline	1.133	0.628	0.785	0.241
Chinese Journal of Chemical Physics	Chemistry	1.316	0.862	0.792	0.024
Chinese Journal of Electronics	Electronic Technology	1.069	0.475	0.598	0.084
Chinese Journal of Natural Medicines	Pharmacy	1.117	0.558	0.622	0.071
Chinese Journal of Oceanology and Limnology	Oceanography & hydrology	1.131	0.675	0.713	0.288
Chinese Journal of Polymer Science	Chemistry	1.285	0.862	0.787	0.204
Chinese Medical Journal	Medicine	1.018	0.626	0.634	0.481
Chinese Optics Letters	Optoelectronics & Laser Technology	1.297	0.889	0.710	0.165
Chinese Physics B	Physics	1.146	0.889	0.678	0.832

Chinese Physics C	Physics	1.333	0.889	0.918	0.074
Chinese Physics Letters	Physics	1.214	0.889	0.770	0.726
Communications in Theoretical Physics	Physics	1.326	0.889	0.838	0.087
Frontiers of Chemical Science and Engineering	Chemical Engineering, Multidicipline	1.179	0.628	0.895	0.017
Frontiers of Computer Science	Computer Science & Technology	1.128	0.537	0.757	0.028
Frontiers of Materials Xcience	Materials Scienc, Multidicipline	1.270	0.784	0.911	0.007
Frontiers of Mathematics in China	Mathematics	1.252	0.774	0.881	0.129
Frontiers of Medicine	Medicine, Multidicipline	1.172	0.626	0.749	0.033
Frontiers of Physics	Physics	1.335	0.889	0.885	0.032
Frontiers of Structural and Civil Engineering	Civil Engineering	1.106	0.473	0.892	0.002
Genomics Proteomics & Bioinformatics	Bilology basic subject	1.330	0.883	0.899	0.051
Geoscience Frontiers	Geosciences	1.180	0.639	0.863	0.056
Insect Science	Entomology and Zoology	1.233	0.777	0.843	0.254
Journal of Arid Land	Geography	1.108	0.490	0.697	0.018
Journal of Bionic Engineering	Mechanical Engineering Design	1.091	0.443	0.794	0.014
Journal of Computer Science and Technology		1.117	0.537	0.748	0.069
Journal of Earth Science	Geosciences	1.136	0.639	0.660	0.151
Journal of Energy Chemistry	Oil and Gas Engineering	1.053	0.341	0.752	0.013
Journal of Environmental Sciences	Environmental Sciences & Resource Sciences	1.109	0.558	0.807	0.199
Journal of Forestry Research	Forestry	1.053	0.414	0.667	0.080
Journal of Genetics and Genomics	Bilology basic subject	1.279	0.883	0.675	0.197
Journal of Geographical Sciences	Geography	1.103	0.490	0.789	0.051
Journal of Hydrodynamics Series B	Water Engineering	1.048	0.363	0.738	0.058
Journal of Integrative Agriculture	Agriculture, Multidicipline	1.033	0.363	0.604	0.066
Journal of Integrative Plant Biology	Botany	1.054	0.742	0.618	0.601
Journal of Iron and Steel Research, International	Metal Materials	1.174	0.708	0.711	0.195

Journal of Materials Science & Technology	Materials Scienc, Multidicipline	1.238	0.784	0.829	0.228
Journal of Molecular Cell Biology	Bilology basic subject	1.327	0.883	0.873	0.073
Journal of Mountain Science	Geography	1.107	0.490	0.759	0.026
Journal of Ocean University of China	Oceanography & hydrology	1.179	0.675	0.730	0.108
Journal of Rare Earths	Materials Scienc, Multidicipline	1.209	0.784	0.772	0.321
Journal of Systematics and Evolution	Botany	1.075	0.742	0.597	0.486
Journal of Wuhan University of Technology Materials Science Edition	Materials Scienc, Multidicipline	1.255	0.784	0.834	0.108
Journal of Zhejiang University Science A	Natural Science(University journals), Multidicipline	1.138	0.596	0.751	0.106
Journal of Zhejiang University Science B	Bilology basic subject	1.324	0.883	0.776	0.054
Light Science & Applications	Physics	1.333	0.889	0.922	0.077
Molecular Plant	Botany	1.218	0.742	0.896	0.311
Nano Research	Materials Scienc, Multidicipline	1.263	0.784	0.928	0.134
National Science Review	Natural Science, Multidicipline	1.198	0.684	0.880	0.127
Neural Regeneration Research	Neurology & Psychiatry	1.267	0.831	0.697	0.123
Neuroscience Bulletin	Neurology & Psychiatry	1.296	0.831	0.858	0.039
Particuology	Chemical Engineering, Multidicipline	1.168	0.628	0.833	0.084
Pedosphere	Pedology	1.077	0.498	0.715	0.136
Petroleum Exploration and Development	Oil and Gas Engineering	0.459	0.341	0.307	1.000
Protein & Cell	Bilology basic subject	1.318	0.883	0.870	0.156
Rare Metals	Metal Materials	1.186	0.708	0.753	0.177
Research in Astronomy and Astrophysics	Astronomy	1.243	0.930	0.824	1.000
Rice Science	Agronomy	1.101	0.474	0.759	0.022
The Crop Journal	Agronomy	1.103	0.474	0.804	0.019
Transactions of Nonferrous Metals Society of China	Metal Materials	0.997	0.708	0.702	1.000
Virologica Sinica	Microbiology & Virology	1.202	0.778	0.663	0.214

Conclusion and future works

First, English as a global lingua franca has affected the language choosing of publishing industry. If Non-native English countries want to show in the stage of world S&T, they have to recognize and make full use of English language. Q-value can reflect and measure the academic communication effect of English-language academic journals published in non-native English countries. This index can also be used as evaluation and monitoring tool of English-language academic journals published in non-native English countries.

Second, as we can see from formula (1), the number of Q-value is the length from 0 to the location of the journal. Journals are in a space which is comprised by prevalence P, centrality C and international diffusivity D. The relatively difference between these 3 indexes can reflect journal's different angle composition.

Third, in some certain discipline, English language is more prevalent. The researchers are likely to use English language articles and journals, thus English-language journals will play a more important role in these areas. For example, in China, the top 10 disciplines in English prevalence index were shown in table 2. Prevalence index can be used to measure the internationalization degree of disciplines and the demand degree for English of disciplines, and also can be used to reflect the potential communication value of English-language journals in disciplines. Such as in China, the communication value is more higher in disciplines of *Astronomy, Physics* and so on.

Table 2. Top 10 disciplines in English prevalence index in China

discipline	*English prevalence index*
Astronomy	0.930
Physics	0.889
Bilology basic subject	0.883
Chemistry	0.862
Endocrinology & Metabology & Rheumatology	0.834
Neurology & Psychiatry	0.831
Haematology & Nephrology	0.820
Internal Medicine, Multidicipline	0.815
Biology Engineering	0.809
Oncology	0.806

Forth, Q-value can evaluate journals in different disciplines because we already take the difference in different disciplines into account.

Fifth, in this study we use the data of English-language journals published in China. There are databases at home like CSTPCD and in international such as JCR. Only in this way, the Q-value can be calculated and it be used as international comparison.

Last but not least important, Q-value cannot replace the traditional indexed such as impact factor, total citations, et al. it can be used as supplements for non-native English countries to evaluate their English-language journals.

Acknowledgments

This research was supported by National Social Science Foundation of China (Project Number:15BTQ059) and Institute of Scientific and Technical Information of China(Project Number:ZD2018-18)

References

Abram de Swaan . (2008). *Words of the World: The Global Language System*. Flowercity Press. Guangzhou.

Beaver D., Rosen R. (1978). Studies in scientific collaboration: Part I: The professional origins of scientific co-authorship. *Scientometrics*, 1(1), 65.

Bennett, K. (2015). Towards an epistemological monoculture: Mechanisms of epistemicide in European researchpublication. In English as a Scientific and Research Language; Pló Alastrué, R., Pérez-Llantada, C., Eds.;Walter de Gruyter: Berlin, Germany; Boston, MA, USA, 9–35.

Burgess, S. (2002). Packed houses and intimate gatherings: Audience and rhetorical structure. In Academic Discourse; Flowerdew(Ed), Pearson Education: London, UK, 196–215.

Crocker MW., Knoeferle P., Mayberry MR. (2010). Situated sentence processing: The coordinated interplay account and a neurobehavioral model. *BRAIN AND LANGUAGE*, 112 (3), 189-201.

Daniel G., Guido S., Gerd A. (2006). The demise of the randomised controlled trial: bibliometric study ofthe German-language health care literature, 1948 to 2004. *BMC Medical Research Methodology*, 6, 30.

EggerE., Zellwegere ZT., Schneider M, et al. (1997). Language bias in randomised controlled trials published in English and German. *LANCET*, 350(9074), 326-329.

Gonza´lez-Alcaide, G., Valderrama-Zuria´n, J. C., & Aleixandre-Benavent, R. (2012). The impact factor in non-english-speaking countries. *Scientometrics*, 92(2), 297–311.

Huh, S. (2014). How Far Has the International Neurourology Journal Progressed Since Its Transformation Into an English Language Journal? *INTERNATIONAL NEUROUROLOGY JOURNAL*, 18(1), 3-9.

ISTIC.(2018). *Chinese English-language S&T journal citation reports*. ISTIC. Beijing

ISTIC.(2018). *Chinese S&T journal citation reports*. Scientific and Technical Documentation Press. Beijing

Liu F., Hu GY., Tang L, et al. (2018). The penalty of containing more non-English articles. *Scientometrics*, 114(1), 359–366.

Liang LM., Rousseau, R., Zhong, Z(2013). Non-English journals and papers in physics and chemistry: bias in citations?. *Scientometrics*,954(1), 333–350.

Ma, Z., Wang N., Zhou GZ., et al.(2012). The analysis of mutual citation network on patterns of Chinese S&T core journal groups. *Studies in Science of Science*, (7), 9–15.

Ma, Z., Pan YT ., (2012). On the communication value of English academic journals published in China based on citation analysis. *Acta Editologica,*24(4), 307–310.

Mintomo Yuasa. (1979). The science activities centre transfer. *Science and philosophy*,2,53-73.

Mongeon, P., Paul-Hus, A. (2016). The journal coverage of Web of Science and Scopus: a comparative analysis. *Scientometrics*,106(1), 2013–228.

Moskaleva, O., Pislyakov, V., Sterligov, I., et al., Russian Index of Science Citation: Overview and review. *SCIENTOMETRICS*, 116(1), 449–462.

Feng Y., Liang LM. (2000). Analysis of the Space- time Characteristic of the Transference of World Scientific Center and Its Subject Arrangement (I), *Science of Science and Management of S.& T*, 5, 4-8.

Narvaez-Berthelemot N. (1995). An index to measure the international collaboration of developing countries based on the participation of national institutions: The case of Latin America. *Scientometrtics*, 34(1), 37-40

Pao, M.L. (1982). Collaboration in Computational Musicology. *Journal of the American Society for Information Science*, 33(1), 38

Solovova, O., Santos, J.V., Verissimo, J., (2018). Publish in English or Perish in Portuguese: Struggles and Constraints on the Semiperiphery. *PUBLICATIONS.* 6(2),25

Van Raan, A. F. J., van Leeuwen, T. N., & Visser, M. S. (2011). Severe language effect in university rankings: Particularly Germany and France are wronged in citation-based rankings. *Scientometrics*, 88(2), 495–498.

Wang SH., Wang HJ., Weldon, P.R. (2007). Bibliometric analysis of English-language academic journals of China and their internationalization. *SCIENTOMETRICS,* 73 (3), 331-343

Yu ZL., Jia J. (2017). An analysis on co-authored status of Chinese Scientific and Technical papers. *Global Science, Technology and Economy Outlook,*32 (11-12), 92-100.

Zhao HZ (1984). *Introduction to Scientific Competence Science*. Beijing: Science China Press.

Zitt M., Perrort F., Barre R. (1998). The Transition from "National" to "Transnational" Model and Related Measures of Countries' Performance. *Journal of the American Society for Information Science*, 49(1),30-42.

Reflections on the Science of Team Science

Yuxian Liu[1], Ronald Rousseau[2], Yishan Wu[3]

[1] *yxliu@tongji.edu.cn*
Yunnan University, School of History and Archives, Cuihubeilu 2, 650091 Kunming (China)
Tongji University, Tongji University Library, Sipinglu 1239, 200092 Shanghai (China)

[2] *ronald.rousseau@kuleuven.be*
KU Leuven, MSI, Facultair Onderzoekscentrum ECOOM,
Naamsestraat 61, 3000 Leuven, Belgium &
University of Antwerp, Faculty of Social Sciences,
Middelheimlaan 1, 2020 Antwerpen, Belgium

[3] *wuyishan@istic.ac.cn*
Chinese Academy of Science and Technology for Development, Yuyuantannanlu 8, 100038 Beijing (China)

Abstract

In this submission we provide a short overview of the Science of Team Science (SciTS). Starting from the notion of a scientific team, we move to interdisciplinary studies and finally the Science of Team Science itself. We describe the main areas of research in this field. As co-authorship networks may grow over time, leading to a future "global brain", it is of the utmost importance to understand how teams work and what leads to their failure or success. Such an investigation needs a collaborative, interdisciplinary and international effort. This will result in reaching the main objective of SciTS, namely to use science to transform the ways researchers actually do science in order to enhance its effectiveness.

1. Introduction

The days of the lone thinker (in an ivory tower?) or the lone inventor (in his backyard?) are gone. Nowadays, most scientific activities are performed in teams. According to (Rey-Rocha et al., 2006) the term 'team' in a scientific context may be defined from two perspectives: an output-based and an input-based. In output-based studies a team is defined based on co-authorship, while in input-based studies a team is based on existing administrative arrangements. Rey-Rocha et al. (2006) prefer the input-based approach but besides belonging to the same administrative unit, they require team members to work on common lines of research, sharing tasks and resources to achieve common objectives. We note that it seems that this notion of a team does not include interdisciplinary teams as members of such teams usually belong to different administrative units. Hence we suggest removing reference to administrative units from the definition of a team, leading to the following definition:

A research team is a collection of people who are organized as a unit, combining their efforts to achieve a common objective: discovering something new through research by sharing information, resources and expertise.

In many cases some (or all) members of a research team will become co-authors on a publication or co-inventors on a patent once their objective is reached. We further note that teams may differ in size, in duration and in composition. Moreover, given the rise of collaborative research and the corresponding increase in the number of team members, it might be better if all members of a team, including e.g., technicians or citizens (in the case of citizen science) become known as contributors of the final outcome, analogous to movie end credits in which all contributors of the movie are mentioned, from the director and leading stars to the set decorators, gaffers and the digital effects computer technicians (Rennie & Yank, 1998; Rousseau et al., 2018, p. 32).

Rey-Rocha et al. (2006) observe that scientists working in well-established teams have an advantage above others, working in un-consolidated groups.

2. What is team science?

Team science is science performed by teams. As a team is not a loosely aggregated bunch of persons, team science naturally implies collaboration. Consequently, a study of teams is a study of collaboration. Teams may differ in many aspects, one of them being their duration. Some teams are temporary, e.g. for one project, while others may last longer, even much longer (Bu et al., 2018). Already more than a decade ago Wuchty et al. wrote an extremely influential paper (Wuchty et al., 2007) in which they analyzed 50 years of research, consisting of more than 20 million scientific publications. They found that teamwork was not only increasingly dominant among all sciences, including the social sciences and humanities, but also in patent applications. Furthermore, they found that team-based publications had generally a higher impact than papers written by individuals. This paper was followed by (Jones et al., 2008) in which the same leading investigators observed that inter-university collaborations increased the concentration of scientific knowledge in a group of elite universities.

3. Up the ante: interdisciplinary research

Braun & Schubert (2003) noted that interdisciplinarity may vary depending on several factors such as:

-) the number of disciplines which are involved
-) the degree of similarity between them
-) the degree of integration
-) the novelty and creativity involved in the combination.

The increase in the number of teams and in their influence on science can be attributed to two diverging phenomena: a growing demand for knowledge requiring greater specialization on the one hand, and a demand for solutions of problems with a broader scope, on the other. Only large interdisciplinary groups can meet these two demands.

Teams active in interdisciplinary science (Morillo et al., 2001; Rousseau et al., 2019; Wagner et al., 2011) consist of members with different expertise related to the disciplines concerned. As mentioned above, such teams are often built to address a broad array of complex and interacting variables.

Policy makers often see interdisciplinary teams as a way to accelerate scientific innovation in a cost-effective way. They also see interdisciplinary teams as a way to translate scientific knowledge into effective policies and practices (Armstrong & Kendall, 2010). Walsh et al. (2019) found that interdisciplinarity is associated with greater retraction rates. This conclusion means that, because of the division of labor, (larger) interdisciplinary teams more often lead to “problematic” results; Walsh et al. (2019) even refer to this situation as “pathogenic”. Complex interactions of multiple and different aspects, hence involving multiple variables, may lead to uncontrollability.

4. The Science of Team Science

4.1 Definition of the field

According to the National Academies of the USA (National Research Council, 2015) the Science of Team Science (SciTS) can be defined as follows:

The Science of Team Science is a new interdisciplinary field that empirically examines the processes by which large and small scientific teams, research centers, and institutes organize, communicate, and conduct research. It is concerned with understanding and managing circumstances that facilitate or hinder the effectiveness of collaborative research, including translational research. This includes understanding how teams connect and collaborate to achieve scientific breakthroughs that would not be attainable by either individual or simply additive efforts.

The term "Science of Team Science" was first introduced in 2006 at a conference called *The Science of Team Science: Assessing the Value of Transdisciplinary Research,* held in Bethesda (MD). Based on the conference's proceedings the emerging SciTS field was further developed in a supplement to the *American Journal of Preventive Medicine*, published in July 2008 (Stokols et al., 2008).

The following aspects are important in the Science of Team Science (Stokols et al., 2008): defining key terminology to facilitate communication between different subject areas; determining methods and models for the study of team science; determining the structure and organization of team science and how teams are formed; team characteristics and dynamics; effective leadership; measuring cognitive distance between team members; team composition and effective communication style. Like with many emerging fields, the delineation of SciTS is still not clear. Yet, SciTS focuses on understanding and enhancing the organization, processes and outcomes of team science. Its goals are: a) to find internal and external factors that maximize the efficiency, productivity, and effectiveness of team science and b) to apply the obtained knowledge.

4.2 Internal factors of a team

4.2.1. Team composition

There may be many internal factors at play, team composition being a fundamental one. Team composition refers to the attributes, e.g., skills, gender, culture, and personality, as distributed over team members. Leadership and team spirit are important team attributes, which may largely determine teamwork effectiveness. Further important attributes are diversity and inclusion, including race, gender, rank, etc. In this context it is worth mentioning that recent investigations have shown that gender diversity may lead to a better use of the expertise of each team member (Joshi, 2014). This is especially true when teams have a large percentage of highly educated women. Gender diversity may also spark new discoveries by broadening the viewpoints, questions, and areas addressed by researchers. By cultivating gender diversity, teams can overcome inherent biases and better reap the full rewards of the team's expertise. Yet, inclusion does not end with gender diversity (Asai, 2019). Inclusion is a feeling of belonging. Any type of diversity be it in terms of gender, sexual preference, nationality, language, age, disciplinary background, basic education, attitude with respect to risk taking or pragmatism, should be embraced and appreciated by the team (Solomon, 2018). Yet, even Solomon (2018) admits that a similar cultural background makes it easier to build trust among members.

The composition of a team shapes the emergence of affective states, behavioral processes, and cognitive states (the ABCs of teamwork), which ultimately affect how teams meet their objectives (Bell et al., 2018). The term ABCs of teamwork refers to the "climate" of a team. The factors in this dimension include two scales: climate scale and attitude scale. The climate scale includes deep diversity (measuring the extent to which the team encourages diversity of ideas, values, and experience), innovation, continual learning, psychological safety (measuring if a team member feels safe to bring up problems and tough issues); the attitude scale includes satisfaction, commitment, citizenship behaviors, withdrawal, psychological empowerment in the workplace, task engagement, work conflict, etc. some fields use these scales to understand how team procedures and policies effect the quality of team output. Questionnaires can be designed to survey the situations of each scale in different teams. Results are then used to help in the development of healthy teams.

4.2.2 Team functioning

Teamwork is the process through which team members collaborate to achieve task goals. More precisely, the term 'teamwork' refers here to the activities through which team inputs translate into team outputs, not only in terms of products, but also in terms of team effectiveness and satisfaction. Hence a study of teams is a study of collaboration. Collaboration is the main concern of the team process. Under the influence of the ABCs of a team, team members collaborate, leading to an emerging team structure. They communicate and share responsibility according to their roles. If they cannot communicate well, the team may fall apart and cannot fulfil its function. If the team members communicate well, they may form a resilient co-authorship network to perform as a global brain. In Börner et al. (2005) the authors illustrate how co-authorship networks grow over time, leading to a future "global brain" similar to the complex pattern of neural connections in the human brain that underlie cognition and behavior. They conclude that only science driven by high-impact co-authorship teams will be able to dynamically respond to increasing demands on information processing and knowledge management.

As teams consist of individuals, also an individual-level analysis is part of the Science of Team Science. In one of the studies related to the individual level, members of teams were queried about their experiences as team members (Börner et al., 2010). Another point is the cost of coordinating large teams, especially those consisting of scientists originating from fields with different research traditions (Cummings & Kiesler, 2007).

4.3 External factors

Acts of collaboration may occur within the team as well as outside the strict boundaries of the team. Such outside factors, which may be quite complicated and not immediately under the control of the team, include external circumstances and environments. This was mentioned by Stokols et al. (2008) who distinguished the following broad areas in the Science of Team Science: studying environmental and contextual factors that influence the effectiveness of team initiatives and developing theoretical models to account for the circumstances and environments under which team science initiatives are more or less effective.

Among other external factors, more attention should be paid to the role of culture. Culture not only plays a role within a team, but also through the external environment influencing team effectiveness. When researchers do not only originate from different fields of science but also from geographically dispersed places or from a different religious background, success of such an international research team requires an understanding of and sensitivity to cultural

differences. When the team is composed by members who are from different countries, scholars have argued that if psychologists are to gain a true understanding of human behavior, culture should be central to research and theory. The research on teams is an area where better integration between the mainstream and cross-cultural literatures is critically needed, given the increasing prevalence of multicultural teams. Guided by accepted frameworks of team effectiveness (Ilgen et al., 2005) and culture (Giorgi et al., 2015), Feitosa et al. (2018) extracted several key assumptions from the mainstream literature on these topics. Through a process of comparing and contrasting, they determined which components of current models are upheld and debunked when seeking to generalize these models to cultural contexts outside North America. These authors provide a foundation for future research taken culture into account, and facilitate a better understanding of human behavior within a team context.

4.4 Institutional factors

Besides the larger region or country and its culture, also the institutional and departmental culture plays a role in teamwork. In which department is the team set up? Does the team have a chance to institutionalize into a normal institution? This point is of no importance when the team is establish for just one project, but may become essential for teams that are meant to be consolidated (Bu et al., 2018). Will a team set up for a longer time have more chances to become a department?

5. Reflections on the measurement of team science and factors involved in this

Researchers specializing in the Science of Team Science have created measures to assess team science processes and outcomes, and to influence contextual and environmental conditions (Klein, 2008). Such measures can be applied to evaluate team science and consequently improve their quality. Their final aim is to develop best practices. Nowadays such best practices include the formation of mixed gender teams.

Many factors can affect team functioning and should be considered when evaluating teams: training differences, social status, differences in responsibility, gender, age and culture, among others. Due to increasing specialization, an increasing number of individuals are only involved in a team on a temporary basis, which complicates collaboration. Not all participants have the same perception of the team structure and its operation. Different specialists have to be aware of their position in the team and pay attention to their legal, ethical and professional responsibilities. Multidisciplinary, well designed training programs are needed (Booij & van Leeuwen, 2008). Unique contextual factors within the scientific enterprise create circumstances to study these teams in context, and provide opportunities to advance understanding of other complex forms of collaboration (Hall et al, 2018).

6. Discussion

When scientists work close together over a long period of time this may lead to conflicts. In the worst case collaboration has to be abandoned. Some groups of people never become a real team with a common goal, and obstacles can become so overwhelming that trying to solve conflicts may be in vain. Tension is an unavoidable part of group dynamics. Already in 1944 Brozek and Keys (1944) wrote that “Cooperative work is a social art and has to be practiced with patience” as quoted by Fiore (2008). In this context Fiore (2008) made the important point that the Science of Team Science must pay special attention to social psychology, organizational psychology and management research, in particular organizational behavior research. Team members and especially team leaders should be educated in training methods developed to overcome the interpersonal, communication, and coordination issues that have been identified in the study of teams (Fiore, 2008; Asai, 2019).

An important focus in the Science of Team Science is to find out if research by interdisciplinary teams lead to better research, in the academic (scientific progress, maybe more citations or more recognition in the form of scientific prizes) and in the societal (larger benefit for society, e.g. in population health) sense of the word, than disciplinary teams. If this is not always the case, under which circumstances or for which type of research problems are interdisciplinary teams more suited?

Some scientific problems are probably more suitable to be solved by a small number of individuals with a similar educational background. For example, a mathematical problem that has not been solved for many years might need an extremely clever and dedicated individual to be solved, the solution of Fermat's last problem by Andrew Wiles springs to mind here (Singh, 1997). Besides studies in fundamental physics and related Big Science topics, most of the pressing societal issues such as problems related to climate change, prevention and control of infectious diseases or cancers, environmental protection and drug addiction, must rely on the collaboration of team members from different backgrounds. In these cases we have no choice but to work in teams.

7. Conclusion

This review summarizes some findings from the SciTS literature, including the value of team science, team composition and its influence on team science performance, formation of science teams, team processes central to effective team functioning, and institutional influences on team science. The interaction between internal and external factors is essential to understand the mechanism of how teams function. The products of a team may be its publications but often also include other products such as influence on policy and, influence on the administrative structure of a university (if the team is a part of a university). Another essential aspect of the Science of Team Science is the study of the translation of findings by research teams to practice and policy, including design and outcomes of training programs to support team science. As this contribution is just a short review we refer the reader to (National Research Council, 2015) in which a thorough overview of the field has been provided.

We encourage colleagues from the bibliometric and informetric community to look out for new research opportunities to further advance SciTS and better inform policies and practices for effective team science.

We conclude by stating that understanding how teams work and what leads to their failure or success, needs a collaborative, interdisciplinary and international effort. This will result in reaching the main objective of SciTS, namely to use science itself to transform the ways researchers do science in order to enhance its effectiveness.

Acknowledgement: This paper was partly supported by grants 17AKS004, and 18KDALD026 and further by the National Natural Science Foundation of China 71173154.

References

Armstrong, A. & Kendall, E. (2010). Translating knowledge into practice and policy: the role of knowledge networks in primary health care. *Health Information Management*, 39(2), 9-17.

Asai, D. (2019). To learn inclusion skills, make it personal. *Nature*, 565(7741), p.537.

Bell, S.T., Brown, S.G., Colaneri, A., & Outland, N. (2018). Team composition and the ABCs of teamwork. *American Psychologist*, 73(4), 349-362

Booij, L. H. D. J., & van Leeuwen, E., (2008). Teamwork and the legal and ethical responsibility of the anaesthetist. *Current Opinion in Anesthesiology*, 21(2), 178-182. DOI: 10.1097/ACO.0b013e3282f5ff87.

Börner, K., Contractor, N., Falk-Krzesinski, H.J., Fiore, S.M., Hall, K.L., Keyton, J., Spring, B., Stokols, D., Trochim, W., & Uzzi, B. (2010). A Multi-Level Systems Perspective for the Science of Team Science. *Science Translational Medicine,* 2(49), 49cm24.

Börner, K., Dall'Asta, L., Ke, W.M., & Vespignani, A. (2005). Studying the emerging global brain: Analyzing and visualizing the impact of co-authorship teams. *Complexity*, 10(4), 57-67.

Braun, T., & Schubert, A. (2003). A quantitative view on the coming of age of interdisciplinarity in the sciences 1980 -1999. *Scientometrics*, 58(1), 183-189.

Brozek, J., & Keys, A. (1944). General aspects of interdisciplinary research in experimental biology. *Science*, 100, 507–512.

Bu, Y., Ding, Y., Liang, XK. Murray, D.S. (2018). Understanding persistent scientific collaboration. *Journal of the Association for Information Science and Technology*, 69(3), 438-448.

Cummings, J. N., & Kiesler, S. (2007). Coordination costs and project outcomes in multi-university collaborations. *Research Policy*, 36(10), 1620-1634.

Feitosa, J., Grossman, R., & Salazar, M. (2018). Debunking key assumptions about teams: The role of culture. *American Psychologist* 73(4), 376-389.

Fiore, S.M. (2008). Interdisciplinarity as teamwork — How the science of teams can inform team science. *Small Group Research,* 39(3), 251–277.

Giorgi, S., Lockwood, C., & Glynn, M.A. (2015). The many faces of culture: Making sense of 30 years of research on culture in organization studies. *The Academy of Management Annals*, 9(1), 1-54.

Hall, K.L., Vogel, A.L., Huang, G.C., Serrano, K.J., Rice, E.L., Tsakraklides, S.P., & Fiore, S.M. (2018). The Science of Team Science: A review of the empirical evidence and research gaps on collaboration in science. *American Psychologist*, 73(4), 532-548.

Ilgen, D.R., Hollenbeck, J.R., Johnson, M. & Jundt, D. (2005). Teams in organizations: from input-process-output models to IMOI models. *Annual Review of Psychology*, 56, 517-543.

Jones, B.F., Wuchty, S., & Uzzi, B. (2008). Multi-university research teams: shifting impact, geography, and stratification in science. *Science*, 322(5905), 1259-1262.

Joshi, A. (2014). By whom and when is women's expertise recognized? The interactive effects of gender and education in science and engineering teams. *Administrative Science Quarterly*, 59(2), 202–239.

Klein, J.T. (2008). Evaluation of interdisciplinary and transdisciplinary research. *American Journal of Preventive Medicine*, 35(2S), S116-S122.

Morillo, F., Bordons, M., & Gomez, A. (2001). An approach to interdisciplinarity through bibliometric indicators. *Scientometrics*, 51(1), 203-222.

National Research Council (2015). *Enhancing the effectiveness of team science.* Washington, DC: The National Academies Press. https://doi.org/10.17226/19007. Introduced in Chinese by Wu Yishan in his blog: http://blog.sciencenet.cn/blog-1557-887422.html

Rennie, D., & Yank, V. (1998). If authors become contributors, everyone would gain, especially the reader. *American Journal of Public Health*, 88(5), 828-830.

Rey-Rocha, J., Garzón-García, B., & Martín-Sempere, M.J. (2006). Scientists' performance and consolidation of research teams in biology and biomedicine at the Spanish Council for Scientific Research. *Scientometrics*, 69(2), 183-212.

Rousseau, R., Egghe, L. & Guns, R. (2018). *Becoming metric-wise: A bibliometric guide for researchers*. Kidlington: Chandos (Elsevier).

Rousseau, R., Hu, XJ., & Zhang, L. (2019). Knowledge integration: its meaning and management. In: (Glänzel, W., Moed, H.F., Schmoch, U., Thelwall, M., Eds.) *Springer Handbook of Science and Technology Indicators*; HB 40, Chap. 3.

Singh, S. (1997). *Fermat's Last Theorem*. London: Fourth Estate.

Stokols, D., Hall, K.L., Taylor, B.K., and Moser, R.P. (2008). The Science of Team Science: Overview of the field and introduction to the supplement. *American Journal of Preventive Medicine*, 35(2S), S77–S89.

Solomon, Y. (2018). 9 Diversity Dimensions That Increase Innovation, and One Reason They May Not.
https://www.innovationexcellence.com/blog/2018/08/28/9-diversity-dimensions-that-increase-innovation-and-one-reason-they-may-not/ translated in Chinese by Wu Yishan in his blog:
http://blog.sciencenet.cn/home.php?mod=space&uid=1557&do=blog&id=1132871

Wagner, C.S., Roessner, J.D., Bobb, K., Klein, J.T., Boyack, K.W., Keyton, J., Rafols, I., & Börner, K. (2011). Approaches to understanding and measuring interdisciplinary scientific research (IDR): A review of the literature. *Journal of Informetrics,* 5(1), 14 –26.

Walsh, J.P., Lee, YN., & Tang, L. (2019). Pathogenic organization in science: Division of labor and retractions. *Research Policy*, 40(2), 444-461.

Wuchty, S., Jones, B.F., & Uzzi, B. (2007). The increasing dominance of teams in production of knowledge. *Science*, 316, 1036–1038.

A Framework to Measure the Impact of Science of a Research Organization

Edgar Schiebel[1], Martin Eichler[2], Robert Kalcik[1], Thomas Scherngell[1], Caroline Wagner[3] and Matthias Weber[1]

[1] *edgar.schiebel@ait.ac.at, robert.kalcik@ait.ac.at, thomas.scherngell@ait.ac.at, matthias.weber@ait.ac.at*
AIT Austrian Institute of Technology GmbH, Donau City Str 1, A-1220 Vienna (Austria)

[2] *martin.eichler@bak-economics.com*
BAK Economics AG, Güterstrasse 82, CH-4053 Basel (Switzerland)

[3] *wagner.911@osu.edu*
Ohio State University, John Glenn College of Public Affairs
Page Hall, 1810 College Road, Columbus, Ohio 43210 (USA)

Abstract

Recently, expectations regarding the benefits of scientific research for social well-being have continually increased. These growing expectations are linked to the highly ambitious (societal) aims of research and innovation policies which are increasingly shifting their focus from the economic impact, for example productivity gains, economic growth or job creation, to a vast field of goals which aim to address grand, socio-political challenges. Public funded research organizations like universities, competence centers, non-academic research institutes must report on the impact of their research and proposals for research projects need an outlook on societal implications. We propose a framework for the measurement of the societal impact of research with a pathway model. It includes 8 categories and 13 dimensions like economic impact, environmental impact and others. The pathways are focused on the mission of the organization and its activities to measurable impacts on innovation and other dimensions. The methodologies followed vary from science mapping, network analysis to counterfactual analysis.

Aknowledgement:

Many thanks to Lutz Bornmann for fruitful discussions

Introduction

The investigation of the structure and dynamics of societal impact of research has gained increasing interest since the early 1990s, both from a scientific, as well as from a management and a policy perspective (for an overview see Bornmann, 2013) With the implementation of the New Public Management agenda in the late 1980s, science was suddenly not only required to regularly evaluate and report on their accomplishments via internal peer review assessment procedures but also to show impact and relevance of their scientific activities, for example by the count of citations and impact factors. In the early 2000s, evidence had to be provided to demonstrate the value of science for society. Meanwhile, outcomes and effects of publicly funded research will be traced for employment, economic value and other societal impacts (Macilwain, 2010).

There is no standard model to comprehensively assess the societal impact of research. Traditionally, studies of impact assessment have examined the economic effects of R&I funding in terms of improving productivity levels, job creation and economic growth (Guellec & Pottelsberghe, 2004) and in terms of the private and societal rates of return (Scherer & Harhoff, 2000) and (Salter & Martin, 2001). An extensive amount of literature is already dealing with

the quantification of measuring environmental impact and territorial transformation. Societal impact is, however, much less explored, essentially because it is difficult to define and to be analyzed. This shift in emphasis is particularly visible in the new mission-oriented framework of EU and national research (and innovation) policies and it poses additional challenges for evaluation and impact assessment practices.

Some of the few initial and new efforts to assess societal impact date from 2011 onwards. These are most notably the public value approach (Bozeman & Sarewitz, 2011) based in the assessment of public values, i.e. the capacity of research to help achieve societal goals; the payback framework as example of a logic model investigating the various stages of an R&I activity (Donovan and Hanney 2011); The Social Impact Assessment Method (SIAMPI project) focusing on 'productive interactions' as exchanges between researchers and stakeholders in which knowledge is co-produced and valued (Spaapen and van Drooge 2011). Some Dutch organizations cooperate in the ERiC project which has set itself the goal of developing methods for societal impact assessment (ERiC 2010). A main result was reflected in the cognizance that an intensive interaction between researchers and societal stakeholders is necessary. This could be personal contacts in joint projects, participation in networks, co-authorships in publications, common exhibitions, etc. and is called productive interaction.

Another initiative is the SIMPATIC project which expanded the existing quantitative general-equilibrium simulation models of NEMESIS and GEM G3 into the assessment of societal inclusion and environmental impact (Karkatsoulis 2016).

The U.K. Research Assessment Exercise (RAE) has been comprehensively evaluating research in the U.K. since the 1980s and is considered the most well-known national evaluation agency. The REF uses case studies to assess universities according to the quality of research outputs, the vitality of the research environment and the wider impact of research. Recently, the seven Research Councils, Innovate UK and the new organisation Research England have been brought together under the umbrella of "UK Research and innovation". To ensure that publicly funded research has academic, economic and societal impact "*...applicants will have to demonstrate the pathways to impact for their research. Careful consideration of Pathways to Impact is an essential component of research proposals and a condition of funding.*" (UK Research and Innovation, 2019)

Particularly relevant is the national French project in the field of agro-food ASIRPA (Socio-Economic Analysis of the Impacts of Public Agricultural Research), carried out by the National Agricultural Research Institute (INRA) and which conducts in-depth long-term case studies aiming at internal learning and accountability (Joly, Gaunand et al. 2015).

Jaffe, 2015 defines impact categories that are challenges for our societies. He implies that such challenges can be influenced by science in a positive way. Consequently, he turns his focus on the dimensions, defines indicators from macro-statistics and postulates that a contribution for the improvement is influenced by the societal impact of science or has been influenced in the past. Jaffe summarises that different types of science impacts are fundamentally non-commensurable, so it is not possible to derive a single composite metric of all research impacts that would be useful for decision purposes. Jaffe of course also argues that some impact categories cannot be monetised. The value of the work is to arrange comprehensively important dimensions and to give preliminary indications for a metric of societal impacts.

These pioneering approaches offer relevant insights and steps into the daunting field of societal impact. However, collectively, they are still in intense discussion and suffer from three gaps in the approach. Firstly, they all tend to use only one specific method in the social sciences (either quantitative or qualitative methods). Secondly, they tend to focus on individual sectors (agro-food only, or environmental R&I only) or on specific pre-given impacts (the societal inclusion

or environmental impacts). Thirdly, taken together, they do hardly work on metrics on causality chains from research to macro-effects. The new U.K. initiative “pathways to impact” is an attempt to integrate potential impacts directly in the research activity of the researchers as a condition of funding.

The main challenge of our work is to contribute to close the gap of remaining question on causalities and that way to elaborate which effects of the work of research organization can be proved. This is not possible for the measurement of the societal impact from the macro perspective. One example for the macro perspective is how a research organization contributes to better air or to reduce the climate warming. One of the reasons is that very often research results tend to need longer time spans to show societal effects in the mentioned dimensions. Another reason is that research is only one contributor element with a high input character in complex systems where many other actors and effects play important roles (see for example MBIE, 2017). We suggest working with the model of pathways to societal impact when analysing the activities of a research organisation by the identification of its research efforts in various disciplines with science mapping and using a network analysis to shed a light on cooperation with companies, public authorities or service providers. Thirdly we propose to apply the counterfactual analysis which aims to distinguish the impact of a research organization from benefits which would have accrued regardless of the intervention. It provides “…a comparison between what actually happened and what would have happened in the absence of the intervention."(White, H., 2006)

Conceptual framework – model of pathways to societal impact

The fundamental assumption in our conceptual framework labelled as pathway impact model for a research organization – is that the fundamental building block for any effects and impact is rooted in the mission and the impact of the activities of a research organization.

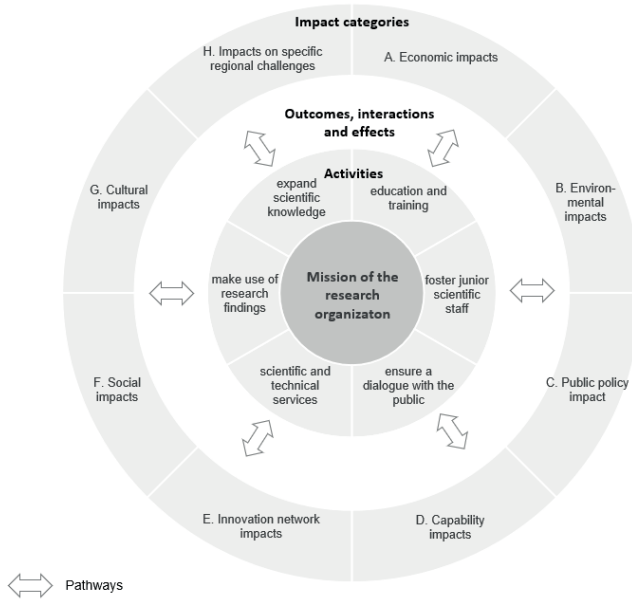


Figure 1: Conceptual framework – Pathway impact model for a research organization

The idea is to identify and describe the pathways that have their origin in results of research, technology transfer, education and training as well as in interactions with the society (see Figure 1).

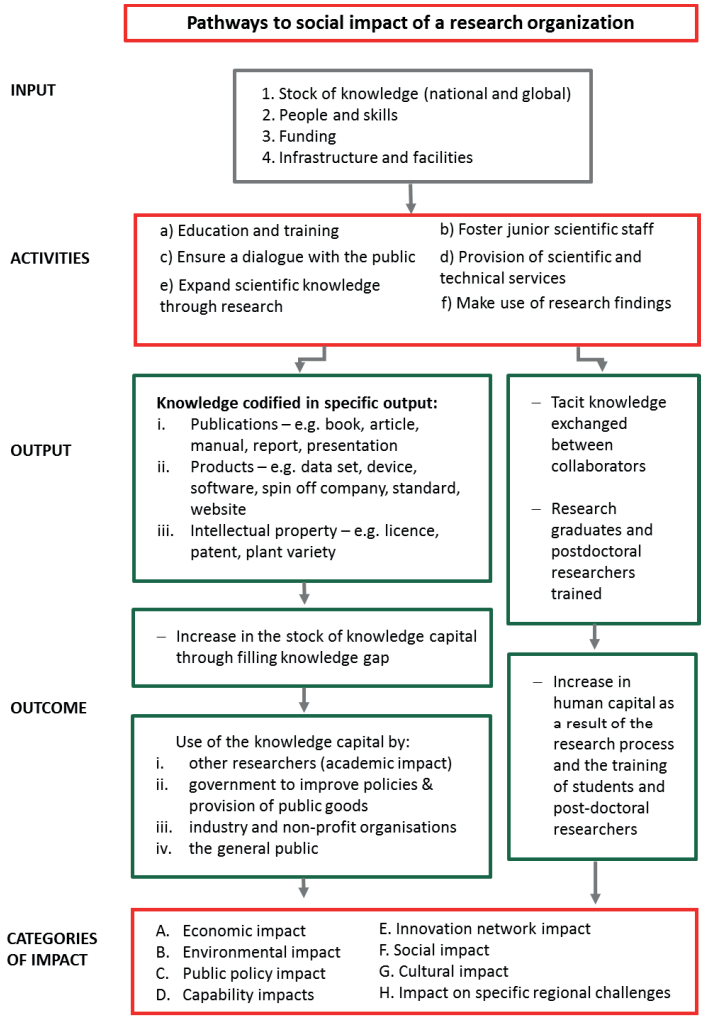


Figure 2: Pathways to societal impact of a research organization

We propose a dynamic pathway model. The centre of the model is the mission of a research organization which is operationalised by several activities of its organisational units. The dimensions of societal impacts form the outer circle are taken from the modified "framework for evaluating the beneficial impacts of publicly funded research" suggested by Jaffe, 2015. The arrows stylise the pathways to societal impacts which could include causalities. That way we take a holistic approach, define pathways and shed light on the causality chains.

Pathways to societal impact are outlined in Figure 2. It is a modification of the generic results chain for science (see MBIE 2017) and includes activities, outputs, outcomes and impact categories for a research organization study.

The inputs consist of the stock of knowledge in the research organization ‘s research fields, people and skills, funding and infrastructure and facilities. The defined purpose of a research organization is summarised in the activities part. Outputs and outcomes are central elements which are supposed to generate effects for societal impact.

The following example for the development of new product or production process and their succeeding economic success roughly illustrates our approach. Let us assume that a research group works on a technology that is essential for the development for a new product or production process. Some results are published, patents are filed, young scientists are educated, accumulated knowledge is further developed, cooperative projects with a company are conducted. The company processes the results, initiates further modifications, can successfully launch the product on the market and has a return on investment what we could measure by for example the share of sales with new products in comparison to a control group of companies that do not cooperate with the research organization and which might be not as innovative and successful. What we must consider is that there are different individual outcomes and economic impacts. Current research about innovation ecosystems clearly shows that the phenomenon is even more complex as it appears in a waterfall model at a first sight. It tells us that even more complex and phenomena exist. That is the reason why we propose to work with network analysis, the structuring and identification of activities with a science and technology mapping procedure and the counterfactual analysis.

Methodology, sources and data

In this paper we focus on the following methodological approaches. Analysis and visualisation (science and technology mapping and collaboration networks) of the structures and dynamics of the research organisations scientific efforts what is the backbone of our framework to study the societal impact. Data sources such as the publication database (Web of Science), project-based research networks of the EU Framework Programs (EUPRO) and patents (PATSTAT) offer basic information for the identification of the thematic output and outcome of the research organization for the pathways to societal impact of research organization.

This analysis enables to identify issues and actors as well as collaboration patterns at different levels (e.g. organisational, regional, rural) and different technological and social domains and subfields. We describe their structure at a local level (e.g. research organization researchers, service organisations and partners of the research organization, alumni and graduates, so to say relevant actors, knowledge ‘gatekeepers’, ‘bridging’ actors), as well as the evolution of these characteristics over time. Turning to a more practical explanation, we will analyse determinants of network structures and dynamics, where we reflect on the following questions. What are thematic issues, how do the different actors engage with each other, what are the drivers/barriers for them or what are impacts of networks on different modes of knowledge output along the pathways to societal impact. By this we can infer the competences in a certain field, the research results, their collaboration partners and target groups of impacts.

Data

In our empirical focus on these analytical approaches developed large-scale datasets for empirically observing R&D collaboration networks are included like Web of Science (publications), PATSTAT (patents) and social media communication (Twitter). One example of a large-scale database is the EUPRO database. It covers EU-FP (FP1-H2020), EUREKA, JTIs (Ecsel, Artemis, Eniac) and COST, comprising information on projects (such as project objectives and achievements, topical allocation, project costs, total funding, start and end date, contract type, information on the call), and participations (standardized name of the participating organisation, organisation type, and geographical location). Another specific asset

of EUPRO is the application of text mining and semantic technologies to attribute projects in a flexible way to relevant topics. As part of the H2020 project KNOWMAK (knowmak.eu), FP projects have been assigned to Societal Grand Challenges (SGC) and Key Enabling Technologies (KETs), and respective subtopics for each KET and SGC. Using these topical assignments of projects, the changing positioning of a research organization in different thematic innovation networks can be measured and illustrated in a systematic way.

Technologies are changing rapidly. New technological fields and novel combinations emerge. This should be reflected in the analysis as well. BAK developed a supplementary patent classification in cooperation with the Swiss Institut für Geistiges Eigentum which aims at the latest developments of cutting-edge technologies and new fields emerging (BAK, 2019). This new technology classification of patents can additionally be included in the technology mapping analysis. The number of patents is increasing substantially in recent years while this is not necessarily true for the commercial and societal value of the individual patent. BAK developed and applies a methodology to assess the value of patents and separate world class patents from the mass of patents. This approach can be used for the technological impact in different dimensions of our pathway model.

Online surveys

Online surveys are useful to gather data in addition to already existing quantitative data. As online surveys are faster, cheaper and more flexible compared to traditional survey methods, specific indicators for certain target groups can be calculated. Those target groups addressed by tailor-made online surveys will be defined according to dimensions and indicators (for example companies collaborating with research organization, graduates from research organization etc.). This method shall on the one hand complement missing data and on the other hand give insights into causality chains. Basically, this method can be applied for all dimensions of the evaluation; however, special emphasis can be put on economic, cultural, knowledge diffusion and environmental issues. The topics for the specific online surveys (2-4 planned surveys) should be tailored to multiple dimensions and directly measurable (perceived/stated) causalities.

Science and technology mapping

Science mapping can be used to thematically structure the research output of a research organization and to identify organisations that directly cooperate with the research organization. It serves for the identification of research results by thematically consistent publication clusters. It delivers technologies, methods and knowledge that have a direct or indirect effect on several impact dimensions that are related to technologies or where research results are helpful, for example in the categories A. Economic Impacts, B. Environmental impacts and others what will describe the pathways from research to several impact dimensions additionally by a semantic analysis of abstracts. When analysing the co-authorships with a network approach. we get information about researchers and their affiliations. Science mapping is a bibliometric approach to analyse and structure a large amount of scientific publications (see e.g. Schiebel E, 2012). The analysis can use peer-reviewed scientific articles recorded in the ISI Web of Science of Clarivate and powered by the AIT software BibTechMon, a comprehensive science mapping tool. The documents will be mapped and clustered by bibliographic coupling of publications (see e.g. Boyack K and Klavans R, 2010). A spring model (see Kopcsa A and Schiebel E 1998) will position similar publications in local virtual regions, and a cluster analysis will lead to a hierarchical structure of similar documents and allows to identify core publications with similar issues. The visualisation of the map (Schiebel, 2012) will occur with a 2D and 3D surface map (see Figure 3).

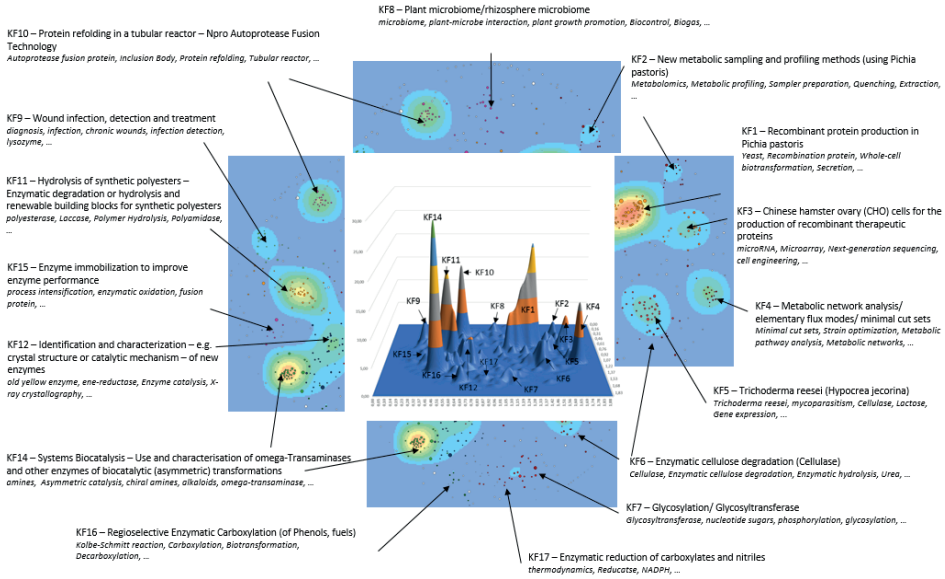


Figure 3: Example for a Science Map of the publication output of the Austrian Centre for Industrial Biotechnology ACIB Austria,

The visualization will indicate together with a hierarchical cluster analysis local agglomerations of research issues that can be relevant for the impact analysis. Additional indicators like the relative publication activity and the knowledge growth factor will help to position a research organization according to its research efforts as competence fields in a research portfolio.

Technology mapping aims at the identification of technologies by patents filed by the research organization and by other actors with a close knowledge flow between the research organization and the applicant. Patents can be structured with the following coupling elements: patent classifications, citations and inventors in the same way as the described the science mapping procedure. Results will be technologies that are related to different impact dimensions and patent indicators will help indicate impacts in dimensions where technologies play an important role (e.g. economic impacts, environmental impacts, public policy impacts, societal impacts, cultural impacts as well as impacts on Swiss specific challenges). A hybrid graph will show technology areas, applicants and inventors with an overlay that indicates the origin (research organization) of the actors.

In a further step, researchers of the competence fields will be matched with inventors of patents where companies are applicants. This essential step will help us in finding companies with close relation to the research organization for which statistical data for the counterfactual analysis can be applied.

Network analysis

To capture research organization innovation network impacts a (social) network analytic (SNA) approach should be employed viewing innovation networks as a graph consisting of nodes (organisations) and edges (joint R&D activities). SNA has come into wide use for the analysis of innovation networks, both in the scientific realms as well as in the policy domain (see Scherngell 2013 for an overview). By means of the SNA approach, the focus to visually illustrate and analyse a research organization collaboration patterns and dynamics should be redirected. Taking a research organization-centric perspective, so-called ego networks can be

examined with the research organization being the central node, inter-linked to a set of other organizations by means of joint projects, joint publications and joint patents with a focus on regional network partners. In doing so, you can characterise the intensity (e.g. number of joint projects) and structure of collaborations (e.g. number of partners, characteristics of partners), disaggregated across thematic fields (e.g. societal challenges) and different organisation types (e.g. firms), and, by this, infer on different kinds of network impacts. Moreover, you can capture to global embedding of the research organization in the wider European innovation network by measuring the development of its network positioning in different topics given by different local and global network centrality measures (see Wasserman and Faust 1994). The global view will illustrate the regions reputation and attractiveness through the network perspective in dimension 13. Of course, network communities with regional organisations can be emphasized.

Clearly for such analyses, a solid and comprehensive empirical backbone is needed. For measuring research organization related R&D collaboration networks, data provided by the research organization on national R&D projects should be used, as well as patent and publication databases. One empirical cornerstone can be the above-mentioned AIT EUPRO database.

Counterfactual analysis

Counterfactual impact evaluation has recently come increasingly into use in structural evaluations (see Bondonio et al. 2016, Janger et al. 2017). In our context, counterfactual analysis aims to distinguish the impact of a research organization from benefits which would have accrued regardless. It is a method of comparison which involves comparing the outcomes between an intervention's beneficiaries (the "treated group") with those of a group similar in all respects to the treatment group (the "control group"), the only difference being that the control group has not been exposed to the treatment, i.e the counterfactual case. The counterfactual position is to imagine an alternative situation where a research organization did not exist and where the activities that they undertake did not take place. Under the umbrella of counterfactual analysis, we can distinguish different statistical methods utilised in different analytical contexts (see Gertler et al 2011). In this kind of impact analysis, most probably two of them should be employed:

Difference-in-difference: Difference-in-difference is an econometric panel data method involving tracking subjects over time. In our context, the method is applied to two groups of companies present in the survey data: those which have engaged in co-patenting or co-publication activities with entities of a research organization and those which have not. The idea behind difference-in-difference is that the time trends in outcomes of the companies that have and have not cooperated with research organization institutions are approximately the same, which is referred to as the common trend assumption. The change in outcomes, e.g. revenues, staff, etc., over time of the companies with the intervention, i.e. research organization cooperation, is compared with the change in outcomes of companies without the intervention. If the common trend assumption is valid, then this is a measure of the intervention's causal effect. Note that there are different strategies to overcome problems such as selection bias or unobserved heterogeneities (see Gertler et al 2011 for details).

Propensity score matching: Propensity score matching sets out to identify pairs of individuals that are comparable in background variables, but where only one received a treatment. Comparing the outcomes of these two individuals yields an estimate of the causal effect of the research organization. A possible application of propensity score matching could be the assessment of capability impacts of the research organization degrees based on data on career paths from LinkedIn. In this case, the propensity score gives the probability of a person's career outcomes, given her background characteristics. Linking pairs of individuals that have a

comparable propensity score, where one received a degree from research organization and the other did not, automatically produces two comparable groups. The advantage of propensity score matching is that it estimates the impact of the intervention, in this case graduation from the research organization, for all individuals where counterfactuals exist. Limitations of the approach are related to non-random allocation of interventions and unobserved covariates that can be avoided by different statistical and other procedures, e.g. in terms of proper sample selection (see Gertler et al 2011 for details).

Example for the metrics for categories and dimensions

An operationalization of the proposed methodology (including the metrics) as well as a thorough explanation why specific indicators are calculated are given for the example of new or improved products or service in the category of economic impacts.

Economic impacts are benefits enjoyed by individual citizens in the form of higher incomes or consumption of higher-quality goods and services. The dimension "New or improved products or services" includes specific innovations sold to the market, and knowledge infrastructure that facilitates innovation or makes existing products better or more valuable.

The proposed metric includes the proportion of companies with product innovation, the share of revenue of new products introduced into the market as a direct measure. As a proxy indicator we propose the proportion of innovative companies with R&D, share of R&D expenditure on revenue. The intermediate outcome can be measured by patent applications, the number of spin-offs from the research organization, licenses taken from the research organization patents, number of collaborative projects with firms. The mentioned indicators on innovation can be calculated from national innovation surveys. When calculating the innovation indicators of firms in the sample with a relation to the research organization. The control group consists of all other firms.

Causal effects of a research organization on firm product innovation are identified via the construction of control and treatment groups among the innovation survey participants in pooled cross-sections or panel data if possible. Firms with a relation to the research organization are identified from co-affiliations in publications, from patents (focusing on product innovations) with co-applicants or inventors from the research organization, from collaborative research projects or spin-offs from the research organization. The matching is used to estimate difference-in-difference coefficients or effects derived from the propensity score matching.

Conclusions

This work proposes a pathway model as a comprehensive and pragmatic framework for the measurement of the societal impact of science for a research organization. It starts with the mission of the organisation, considers the activities of the research organisation und structures them by several dimensions of impact categories. A core idea is that some of the categories are covered by research results like technologies that should have a positive impact on environmental issues, on industrial product and process innovations or the improvement of medical and pharmaceutical treatments. Of course, it depends on the specific kind of research results whether a direct causal effect can be measured or not.

References

Bondonio, D., Biagi, F. and Stancik, J. (2016). Counterfactual Impact Evaluation of Public Funding of Innovation, Investment and R&D. Joint Research Centre. JRC Technical Report EUR 27931 EN; doi:10.2791/240240.

Bornmann, L. (2013). What is societal impact of research and how can it be assessed? A literature survey, Journal of the American Society for Information Science and Technology, 64(2):217-233.

Boyak K. and Klavans, R. (2010): Co-Citation Analysis, Bibliographic Coupling, and Direct Citation: Which Citation Approach Represents the Research Front Most Accurately? Journal of the American Society for Information Science and Technology, 61(12):2389-2404, DOI: 10.1002/asi.21419

Bozeman, B. and Sarewitz D. (2011). Public Value Mapping and Science Policy Evaluation. Minerva 49(1): 1-23.

Donovan, C. and Hanney, S. (2011). The 'Payback Framework' explained, Research Evaluation 20(3):181-183

ERiC (2010). Evaluating the societal relevance of academic research: a guide. Delft, The Netherlands. Delft University of Technology.

ETER (2018). The first comprehensive database on European Higher Education Institutions. https://www.eter-project.com/

Gertler, P. Martinez, S., Premand, P., Rawlings, L., Vermeersch, C. (2011). Impact evaluation in practice. World Bank, Washington DC.

Guellec, D. and Van Pottelsberghe de la Potterie, B. (2004). "From R&D to Productivity Growth: Do the Institutional Settings and the Source of Funds of R&D Matter?" Oxford Bulletin of Economics and Statistics 66(3): 353-378.

Jaffe, A. (2015), A Framework for Evaluating the Beneficial Impacts of Publicly Funded Research, Motu Note No. 15, Motu Economic and Public Policy Research, Wellington, New Zealand

Janger, J., Firgo, M., Hofmann, K., Kügler, A., Strauss, A., Streicher, G., Pechar, H. (2017). Wirtschaftliche und gesellschaftliche Effekte von Universitäten. WIFO, Österreichisches Institut für Wirtschaftsforschung, Wien, Motu, Note #15.

Joly, P.-B., Gaunand A., Colinet, L., Larédo, P., Lemarié, S. and Matt, M. (2015). ASIRPA: A comprehensive theory-based approach to assessing the societal impacts of a research organization. Research Evaluation 24(4): 440-453.

Karkatsoulis, P., Capros, P., Fragkos, P., Paroussos, L. and Tsani, S. (2016). "First-mover advantages of the European Union's climate change mitigation strategy." International Journal of Energy Research 40(6): 814-830.

Kopcsa A. and Schiebel E. (1998). Science and Technology Mapping: A New Iteration Model for Representing Multidimensional Relationships, JASIS; 49, 7-17.

Martin, B. R. (2011). The Research Excellence Framework and the „impact agenda“: Are we creating a Frankenstein monster? Research Evaluation, 20(3) 247-254.

MBIE 2017, The impact of science, Discussion paper, Ministry of Business, Innovation and Employment, New Zealand

Macilwain, C. (2010). What science is really worth. Nature, 465(7299), 682-684.

Salter, A. J. and B. R. Martin (2001). The economic benefits of publicly funded basic research: a critical review. Research Policy 30(3): 509-532.

Scherer, F. M. and Harhoff, D. (2000). Technology policy for a world of skew-distributed outcomes. Research Policy 29(4–5): 559-566.

Scherngell, T. (ed.) (2013). The Geography of Networks and R&D Collaborations. Springer-Physica Verlag, Berlin-Heidelberg-New York.

Schiebel, E. (2012). Visualization of Research Fronts and Knowledge Bases by Three-Dimensional Areal Densities of Bibliographically Coupled Publications and Co-Citations. Special Issue Scientometrics 91, 557-566.

Spaapen, J. and van Drooge, L. (2011). "Introducing ‘productive interactions’ in social impact assessment." Research Evaluation 20(3): 211-218.

Wasserman, S., & Faust, K. (1994). Social network analysis: Methods and applications (Vol. 8). Cambridge university press.

White, H. (2006): Impact Evaluation: The Experience of the Independent Evaluation Group of the World Bank, World Bank, Washington, D.C., p. 3

Upgrading from 3G to 5G: Topic evolution and persistence among scientists

Wencan Tian, Zhigang Hu and Xianwen Wang

tianwen@mail.dlut.edu.cn, huzhigang@dlut.edu.cn, xianwenwang@dlut.edu.cn
WISE Lab, Dalian University of Technology, Dalian, 116024 (China)

Abstract

In this study, we choose wireless mobile telecommunications technology as the research object. We examined the 3G, 4G and 5G literature on IEEE Xplore Digital Library, and analyse the technology upgrading influenced by the topic evolution and other factors. Employing LLR algorithm, the knowledge concept and the expansion of researchers over time in the process of technology upgrading are discussed in depth from the micro level, and the new topics and researchers in each process are studied, topics and researchers with long-lasting vitality are analysed. From the macro level, the phenomenon of technology center migration brought about by technology upgrading is explored. The results of the study show that technology upgrading leads to the emerging of peaks of the knowledge concept, which can lead to a surge in researchers and prolong the life cycle of a field. New topics on the terminology and technical methods, and those long-lasting topics are mostly regarding theories and models; scientists who are good at adjusting or transferring research interests have higher volume and influence than scientists who focus on one field; technology upgrading has created a shift in technology centers that can occur between countries and between institutions within a country.

Introduction

If we would like to define a keyword for future technology, "communication" may be our first choice. With the emergence and popularity of concepts such as autonomous driving and artificial intelligence, the Internet of Everything has become the future development trend, and 5G technology plays an important role in the middle. 1G-5G is a living evolutionary history of technology, and since the beginning of 3G technology, it has really begun to lead a revolution in the mobile Internet. Many scholars have studied the impact of technology upgrading on an industry. For example, Autor (2003) studied how technology upgrading in the computer field change job skills requirements, revealing the positive correlation between technology upgrading and education upgrading. Hung (2014) studied lithium iron phosphate (LFP) batteries in the process of technological change through extensions based on citation-based main path analysis the knowledge diffusion path, revealing that LFP battery technology has completed two complete technical cycles and is in the middle of the third cycle. Funk, (2016) outlined the network approach to research technological change. Karali (2017) examined the impact of energy savings on the US steel industry through analog technology upgrading, indicating that technology upgrading will reduce technology investment costs and save energy. Many scholars have also studied the impact of the technology upgrading process on enterprises (Mainga, Hirschsohn, & Shakantu, 2009; Bustos, 2011; Ho, Ruan, Hang, & Wong, 2016). There are few studies about the technology evolution during the technology upgrading process. In this research, our research questions are, what are the traits during the technology upgrading process from one generation to the next generation? what factors influence the technology upgrading process? Taking wireless mobile telecommunications technology as an example, we hope to answer the research questions from both micro and macro levels.

Data sources and research methods

IEEE Xplore is an academic literature database that provides indexing, summarization, and full-text downloading of literature (including journal articles, conference proceedings, technical standards, and related materials) in computer science, electrical engineering, and electronics. It basically covers the literature of the Institute of Electrical and Electronics Engineers (IEEE)

and the Institute of Engineering and Technology (IET). In this paper, we retrieved 15414, 3957 and 8562 items for "3G mobile communication", "4G mobile communication" and "5G mobile communication" correspondingly. We use the IEEE Term to conduct the search to ensure the precision and recall. In order to solve the author's disambiguation problem, we extract each author's id number and the corresponding article id number based on the JSON data of each article returned by the IEEE Xplore API. Due to the lack of some authors' id numbers and some authors' early papers did not indicate the corresponding id number, we divided the authors into two groups with id and without id, and then compared the full name of each author with the institution. If the two are consistent, we think they are the same author, so that the authors without id are merged into the corresponding id author, and we have deleted the authors without id who are not matched in the whole dataset. Due to the large number of authors, we have done a lot of work on the issue of author disambiguation. The author data obtained in the final processing is shown in the following Table 1.

Table 1. Final processing of author data

Generation	*Total*	*id*	*id Proportion*	*no_id*	*no_id Proportion*
3G	31,074	26,611	85.64%	4,463	14.36%
4G	10,524	8,597	81.69%	1,927	18.31%
5G	20,953	17,496	83.50%	3,457	16.50%

At the topic level, due to the lack of a large number of article keywords, we use the IEEE Term that the IEEE database give to each article. We processed the keywords and obtained the distribution of 3,479 keywords of the dataset (total word frequency 204,481, 3G word frequency 124,934, 4G word frequency 27,772 and 5G word frequency 51,775). In order to study the self-growth process from 3G to 4G and to 5G, we first constructed the corresponding papers-authors_id database. Then we use the LLR algorithm to classify 3,479 keywords into the 3G/4G/5G domain. The LLR algorithm can eliminate the absolute number of differences and classify keywords that appear in multiple fields into one area. According to formula (1), we can calculate the expected frequency of each keyword, and then formula (2) can classify each keyword into a field. Formula (3) can calculate the LLR value of each keyword. Finally, the processed data is imported into VOSviewer (Waltman, Van Eck, & Noyons, 2010) for visualization analysis, and the research topics in 3G/4G/5G and the research topics that have remained unchanged in the process of technology upgrading are displayed.

$$(Expected\ Freq)_{ij} = Freq_i \frac{\sum_i Freq_{ij}}{\sum_i Freq_i} \quad (1)$$

$$Group_{ij} = \max[\frac{Freq_{ij}}{(Expected\ Freq)_{ij}}] \quad (2)$$

$$LLR_j = 3\sum_i Freq_i * log \frac{Freq_i}{(Expected\ Freq)_{ij}} \quad (3)$$

i: stands for the 3G/4G/5G field and has a value of 3, 4, 5.
j: represents 3,479 keywords, with values of 1, 2, 3, ..., 3478, 3479.

For the author, we did the same processing as the keyword. However, due to the large number of authors, it take more time to process the data.

Results

Growth of Knowledge concept and researchers

A process of technology upgrading in a field is a process of self-growth that can be reflected in the growth of knowledge concepts and researchers. In this section, the growth rates of the keywords and authors are measured. For example, there are three articles in the 3G field: paper

1, paper 2, and paper 3. Among them, paper1 and paper 2 were published in 2017, and paper 3 was published in 2018. There are two articles in the 4G field: paper4 and paper 5. Paper 4 and paper 5 were published in 2018. Paper1 contains keywords (authors) A, B, C, paper 2 contains keywords (authors) A, C, D, paper 3 contains keywords (authors) A, D, E, paper 4 contains keywords (authors) A, B, F, paper 5 contains keywords (authors) A, F, G. Therefore, there are four new knowledge concepts/researchers in the 3G field in 2017 (A, B, C, D); one new knowledge concept/researcher in the 3G field in 2018(E); and two new knowledge concepts/researchers in the 4G field in 2018(F, G). As shown in Figure 1.

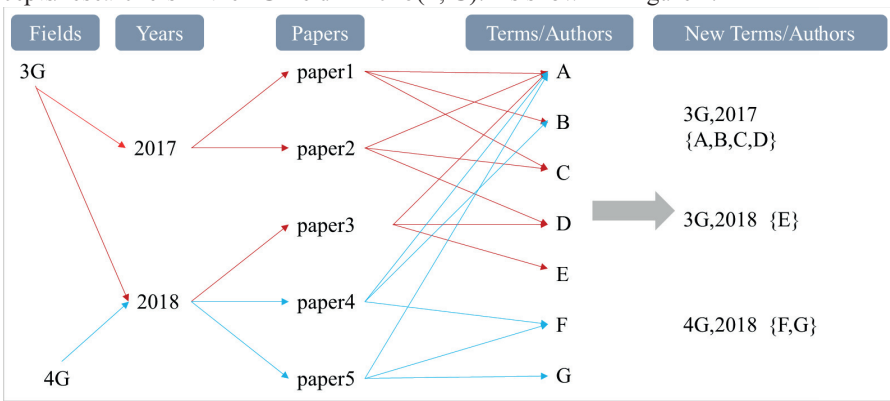


Figure 1. Schematic diagram of the newly added knowledge concepts and researchers

Figure 2 and Figure 3 show the expansion of the newly added knowledge concepts/researchers with the year during the 3G-5G technology upgrading process. The notation on the graph represents the first knowledge concept/researcher in each field. For example, the first knowledge concepts in the 3G field are “Mobile communication”, “Macrocell networks”, etc.; the first knowledge concepts in the 4G field are “Agricultural engineering”, “Laser stability”, etc.; the first knowledge concept in the 5G field was in 2006, only one, which is “C#languages”. After eight years, in 2014, three more emerge, namely “Radio access network”, “System level design and analysis”, “Information exchange”. In Figure 3, since the number of new researchers in the 5G field was 2,643, the number is too large, so no information is added to the newly added personnel. In addition, the first new knowledge concept in the 5G field is in 2006, but the first new researchers emerge in 2015. This is because the 2006 article is published in the 5G field, but the author is previously in 3G field.

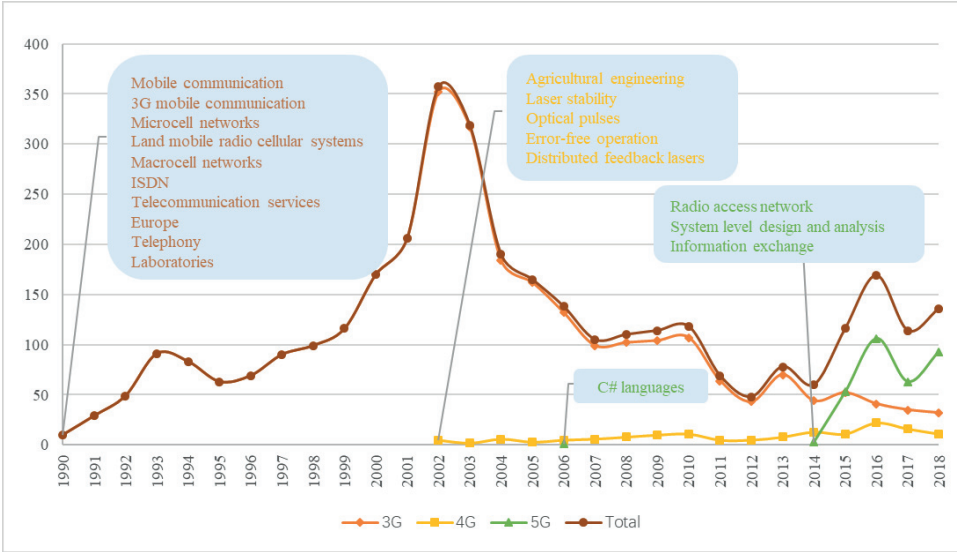


Figure 2. Temporal extension map of new knowledge concepts during the 3G-5G technology upgrading process

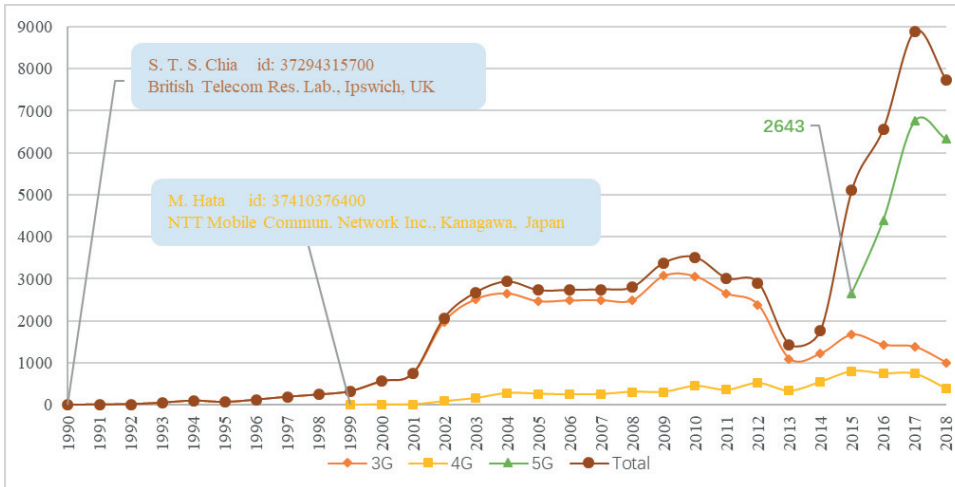


Figure 3. Temporal extension map of new researchers during the 3G-5G technology upgrading process

It can be found that the new knowledge concepts in the 3G and 4G fields are decreasing year by year, and the 5G field has reached a steady state. During the extension process of the overall knowledge concepts, there are two peaks, which are caused by the emergence of 4G and 5G technologies. The technology upgrading brings about a rapid growth of knowledge concepts in a field, thereby extending the life cycle of a field. Figure 3 shows that with the technology upgrading, there is a possibility of explosive growth of researchers. With the advent of the 5G era, the number of new researchers in the 5G field has increased from 0 to 2,643, and has increased at an average annual rate of 1,226 researchers.

Intergenerational characteristics: 3G/4G/5G Terms（3G/4G/5G Authors）

The intergenerational change of mobile communications has lasted for a decade. Since the 3G era, mobile communication has entered a new era. The 3G-5G has undergone earth-shaking technology upgrading. It is a living evolutionary history of science and technology. It has become a symbol of the era of division, and also represents the game history of the world's powers in the scientific and technology revolution and industry transformation. In the 3G era, the situation is dominated by four technology standards . In 2000, the International Telecommunication Union (ITU) identified WCDMA in Europe, CDMA2000 and WiMAX in the United States, and TD-SCDMA in China as the 3G mainstream standard. Around 2005, communication standards ushered in another global competition. In 2010, the European-led FDD-LTE and China-led TD-LTE became the 4G standard, while the US-led WiMAX suddenly retired. At present, the world's powers are all deploying 5G technology globally, and it is expected to be fully commercialized by 2020.

Figures 4 and 5 use the LLR algorithm to classify topics /authors into the 3G, 4G and 5G field. The size of the node represents the word frequency (the number of documents), that is, the larger the node, the larger the word frequency (the number of documents). Three different colours from top left to bottom right (blue for 3G, green for 4G and red for 5G) represent the 3G/4G/5G research themes /researchers, which we call 3G/4G/5G Terms or 3G/4G/5G Authors respectively.

The topics that need to be paid attention to in the 3G field are “multiaccess communication”, “downlink”, “throughput”, and the like. In the 4G field, topics such as “mobile communication”, “bandwidth”, “wireless communication”, “mimo”, “ofdm”, “long term evolution”, and “wireless networks” are important. Among them, OFDM (Orthogonal Frequency Division Multiplexing) technology has good anti-interference performance and is the core technology of 4G. In the 5G field, topics such as “Interference”, “resource management”, and “signal to noise ratio” are indispensable. It is worth mentioning that “cloud computing”, “robotics”, “artificial intelligence”, “smart devices”, and “smart home” are also core research topics in the 5G field, including a large number of studies on “millimeter waves”.

For the authors, their research interests are concentrated on a certain point. For example, 3G author M. Rupp studied 3G word frequency for 248 times, but he studied 4G word frequency only 79 times, and 5G word frequency 73 times.

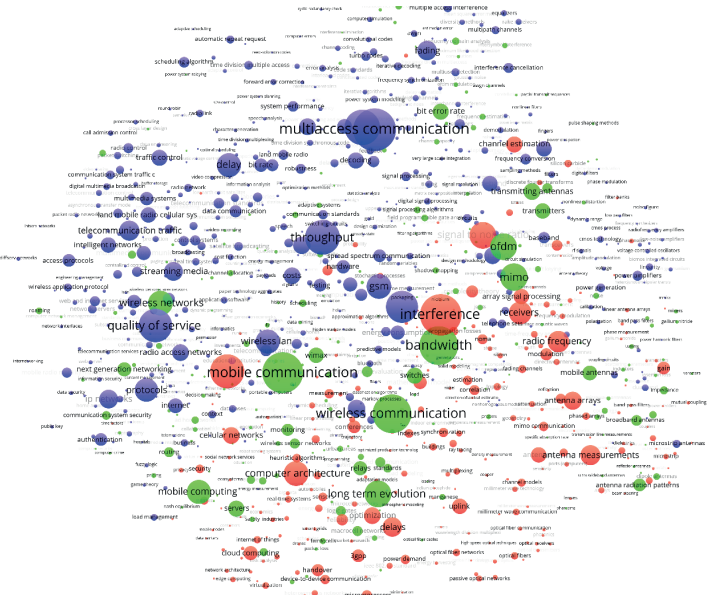


Figure 4. 3G/4G/5G Terms

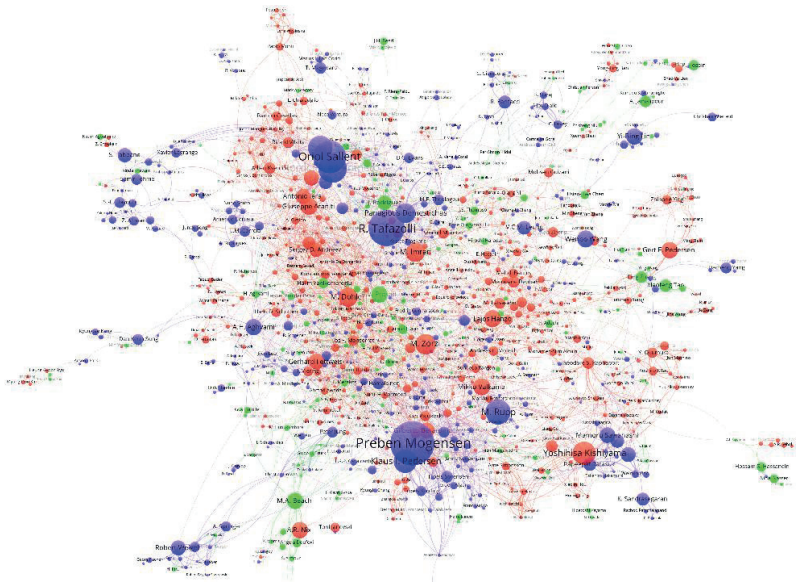


Figure 5. 3G/4G/5G Authors

Intergenerational characteristics: Invariable Terms/Authors

The size of the node in Fig.6 and Fig.7 represents the word frequency or number of documents, that is, the larger the node, the more the number of occurrences (the number of documents). The depth of the colour represents the LLR value. The darker the colour, the smaller the LLR value and the smaller the discrimination (for the visual effect, we have processed the LLR value here: $LLR_j = Max\{LLR_j\} - LLR_j$). We refer to this type of research topic (researcher) with

small LLR values and large word frequency as Invariable Terms (Invariable Authors), which represent the research topics (researchers) that have been inherited in the process of technology upgrading.

As shown in Figure 6, the larger red node represents the Invariable Terms. Research topics such as radio frequency, transmitting antennas, and cellular networks have been running through the 3G-5G technology upgrading process. Further observations revealed that most of these research topics were related to models and algorithms, such as “markov processes” (96 times), “genetic algorithms” (92 times), and “artificial neural networks” (78 times).

As shown in Figure 7, the larger red node represents Invariable Authors. Such authors do not focus their research interests at a fixed point, but with the renewal of technology, they also adjust their research direction. For example, Preben Mogensen of Aalborg University in Denmark studied “downlink” (49 times), “throughput” (38 times) and “multiaccess communication” (26 times) in 3G research; and studied “Long Term Evolution” (30 times) and “bandwidth” in 4G field (28 times), “Multiple-Input Multiple-Output” (21 times); studied “interference” (40 times) and “signal to noise ratio” (28 times) in the 5G field.

Further comparing invariable authors among all Authors, we found that invariable authors have higher volume, influence, and length of study (Table 2).

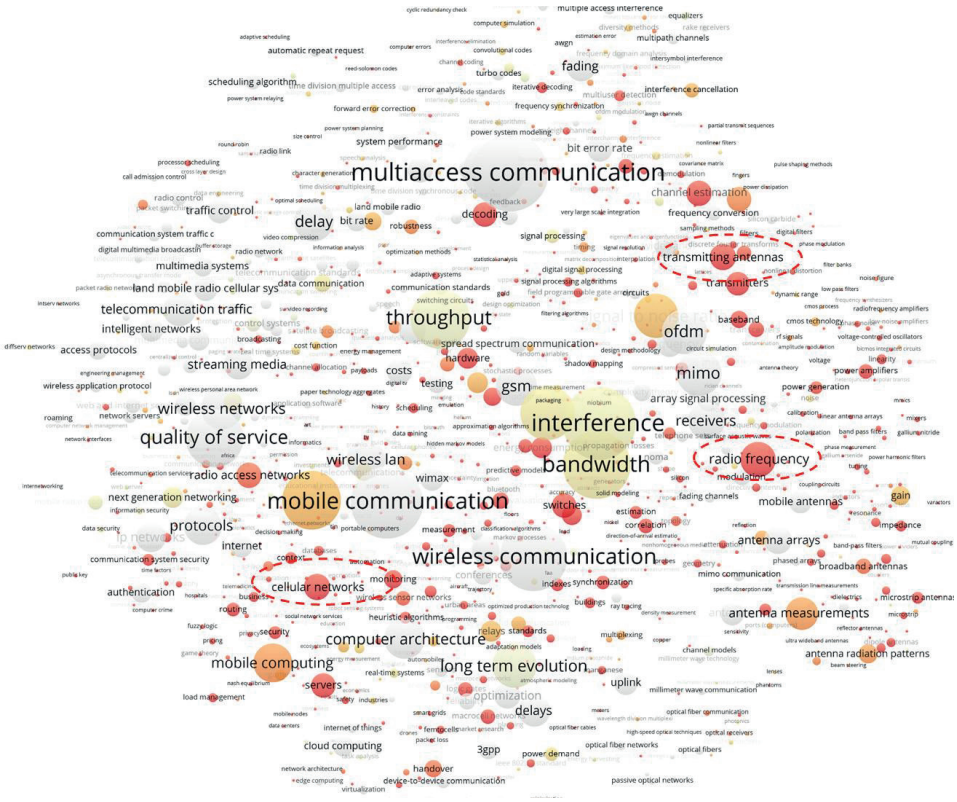


Figure 6. Invariable Terms

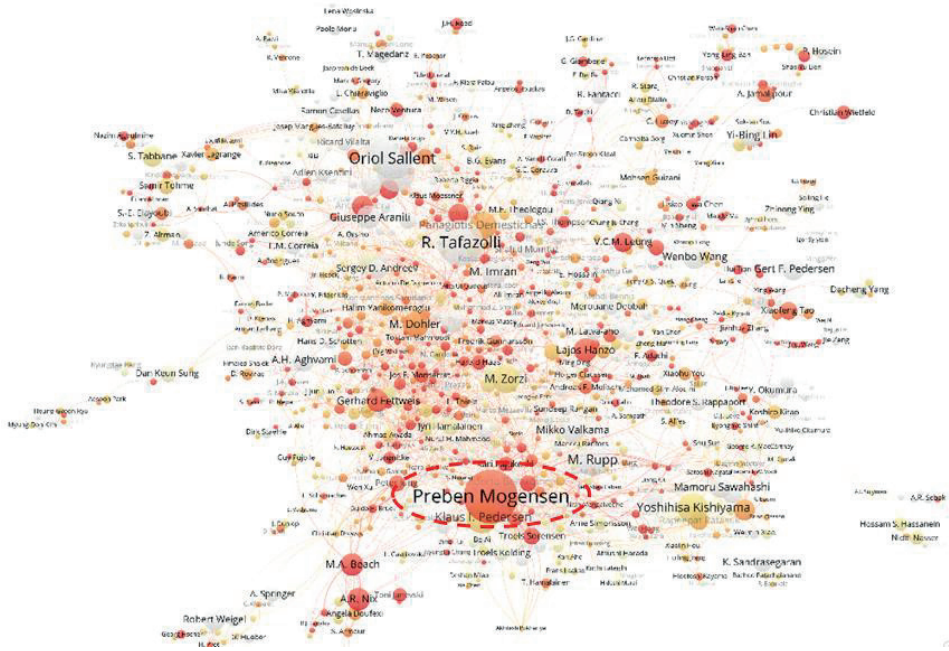


Figure 7. Invariable Authors

Intergenerational characteristics: Technology Center migration

Table 2 lists the top 10 organizations of 3G/4G/5G/ invariable authors, and finds that the migration of technology centers generated by technology upgrading is not only reflected by the migration among countries, but also in the migration of institutions within a country. Like the concept of knowledge and researchers, there are also new births and apoptosis in the technology centers. For example, the Vienna University of Technology (Austria), which was in the leading position in the 3G era, disappeared in the 4G and 5G eras. China's mobile communication technolgycenter has gradually transferred from Beijing University of Posts and Telecommunications to Huawei Technol. Co. Ltd. In contrast, the technology center in Japan has always been NTT DoCoMo.

Table 2. 3G/4G/5G/Invariable Authors (Top10)

Field	*Author*	*id*	*Article quantity*	*Total citations*	*Year of issue*	*institution*
3G Authors	M. Rupp	37276736700	67	756	2001-2018	Vienna University of Technology, Austria
	Wenbo Wang	37280459100	46	770	2001-2012	Beijing University of Posts and Telecommunications, China
	A.H. Aghvami	37271698200	44	388	1992-2017	King's College London, UK
	Yi-Bing Lin	37279533300	43	571	2001-2014	Chiao Tung Univ., Taiwan, China
	Mamoru Sawahashi	37269819300	42	575	2002-2015	NTT DoCoMo, Japan
	Robert Weigel	37266288900	40	142	1999-2018	University of Erlangen-Nuremberg, Germany
	K. Sandrasegaran	37295739900	35	267	2004-2017	University of Technology Sydney, Australia
	A. Springer	37269659900	28	103	1999-2018	Johannes Kepler University of Linz, Austria
	L. Maurer	37298722000	27	110	2000-2010	DICE GmbH&Co KG, Austria
	Claudio Rosa	37294529900	21	494	2003-2016	Aalborg University, Denmark
4G Authors	R.V. Prasad	37275200300	44	684	1993-2015	Delft University of Technology, Netherlands
	Ping Zhang	37274503400	32	235	2001-2011	Beijing University of Posts and Telecommunications, China
	P. Hosein	37269132200	28	57	2002-2010	Huawei Technol. Co. Ltd.(USA)

	Leonardo Goratti	37595187400	22	101	2014-2018	CREATE-NET Research Centre, Italy
	Rose Qingyang Hu	38186830700	19	223	2011-2017	Utah State University, USA
	S. Frattasi	37565541300	11	322	2005-2009	Aalborg University, Denmark
	G. Femenias	37266545400	10	48	1999-2018	University of the Balearic Islands, Spain
	N. Umeda	37278628600	9	81	2000-2012	NTT DoCoMo, Japan
	Frédéric Gianesello	37293985700	9	9	2013-2017	Silicon Technology Development, France
	Farooq Khan	37280820600	8	1011	2006-2012	Samsung Electronics(USA)
5G Authors	M. Imran	37265529100	50	603	2004-2018	King Saud University, KSA
	Tarik Taleb	37265763700	41	966	2009-2018	Aalto University, Finland
	Ricard Vilalta	37966844900	29	34	2015-2018	CTTC/CERCA, Spain
	Zhiguo Ding	37278653500	28	1178	2015-2018	Lancaster University, UK.
	Y. Okumura	37355635900	28	51	1996-2018	NTT DoCoMo, Japan
	Ramon Casellas	37319464800	26	36	2015-2018	CTTC/CERCA, Spain
	S. Suyama	37297955700	24	77	2011-2018	NTT DoCoMo, Japan
	Kari Leppanen	37902704300	21	107	2015-2018	Huawei Technologies (Sweden)
	Paolo Monti	37269380100	21	65	2015-2018	KTH Royal Institute of Technology, Sweden
	Mário Costa	37085451889	19	133	2015-2018	Huawei Technologies Oy (Finland)
Invariable Authors	Preben Mogensen	37268072300	117	1866	1996-2018	Aalborg University, Denmark
	M. Dohler	37273439700	51	1865	2002-2018	King's College London, UK
	Lajos Hanzo	37276255400	45	869	1994-2018	University of Southampton, UK.
	A.R. Nix	37276828100	44	397	1999-2018	University of Bristol, UK
	M.A. Beach	37270115800	44	200	1993-2018	University of Bristol, UK
	Gerhard Fettweis	37271859700	42	474	2000-2018	Technical University Dresden, Germany
	Antonio Iera	38509375800	40	181	1995-2018	University Mediterranea of Reggio Calabria, Italy
	J. Rodriguez	37404807200	38	160	2003-2018	University of Aveiro, Portugal
	V.C.M. Leung	37279114300	36	748	2001-2018	University of British Columbia, Canada
	Antonella Molinaro	37268951900	36	254	1995-2018	University Mediterranea of Reggio Calabria, Italy

Conclusions

During the technology upgrading process, firstly, when the contemporary is alternating, the knowledge concept will have a growth peak, and researchers will explode. Based on this, the technology upgrading extends the life cycle of a field. On the contrary, it is the emergence of those era nouns and the addition of new researchers that in turn promote technology upgrading in one area.

Secondly, researchers' research on a field focuses more on conceptual nouns and technical methods that emerge from the development of the times, and those long-lasting research themes focus more on theory and models. Scientists who are good at adjusting or transferring research interests have a higher volume and influence, which may give us implications that researchers need to adjust research interests or research roadmap to adapt to the time.

Thirdly, the technology center will also migrate. This migration does not only occur among countries, but also among research institutions within a country.

There are some limitations about this study. Although we identifies the expansion of researchers over time, it does not consider the participation of junior researchers and the transfer of mature researchers. We hope to consider these factors in the future research.

References

Autor, DH., Levy, F., & Murnane, RJ. (2003). The skill content of recent technological change: An empirical exploration. *The Quarterly Journal of Economics*, 118(4), 1279–1333. https://doi.org/10.1162/003355303322552801

Hung, SC., Liu, JS., Lu, LYY., & Tseng, YC. (2014). Technological change in lithium iron phosphate battery: the key-route main path analysis. *Scientometrics*, 100(1), 97-120.
Funk, Russell J. , & Owen-Smith, Jason. (2016). A Dynamic Network Measure of Technological Change. *Management Science*, 63(3), 791-817. https://doi.org/10.1287/mnsc.2015.2366
Karali, N., Park, WY., & McNeil, M. (2017). Modeling technological change and its impact on energy savings in the U.S. iron and steel sector. *Applied Energy*, 202, 447-458. https://doi.org/10.1016/j.apenergy.2017.05.173
Mainga, W., Hirschsohn, P., & Shakantu, W. (2009). An exploratory review of the relationship between enterprise training and technology upgrading: evidence from South African manufacturing firms. *International Journal Of Human Resource Management*, 20(9), 1879-1895. https://doi.org/10.1080/09585190903142340
Bustos, P. (2011). Trade Liberalization, Exports, and Technology Upgrading: Evidence on the Impact of MERCOSUR on Argentinian Firms. *American Economic Review*, 101(1), 304-340.
Ho, YP., Ruan, Y., Hang, CC., & Wong, PK. (2016). Technology upgrading of Small-and-Medium-sized Enterprises (SMEs) through a manpower secondment strategy A mixed-methods study of Singapore's T-Up program. *Technovation*, 57-58, 21-29. https://doi.org/10.1016/j.technovation.2016.07.001
Waltman, L., Van Eck, N.J., & Noyons, E.C.M. (2010). A unified approach to mapping and clustering of bibliometric networks. *Journal of Informetrics*, 4(4), 629-635. https://doi.org/10.1016/j.joi.2010.07.002

Burst diffusion of highly retweeted scholarly articles in Social Media

Yunxue Cui[1], Xiaoke Xu[2], Renmeng Cao[2], Zhichao Fang[3], Jianyun Zhou[2], Xianwen Wang[1]

[1]*zmxiaohai@mail.dlut.edu.cn*
[1.]WISE Lab, Faculty of Humanities and Social Sciences, Dalian University of Technology, Dalian 116085, China.
[2]College of Information and Communication Engineering, Dalian Nationalities University, Dalian 116600, China
[3]Centre for Science and Technology Studies (CWTS), Leiden University, Wassenaarseweg 62A, Leiden, 2333 , AL (The Netherlands)

Abstract

In this study, we aim to explore the social media diffusing mechanism of highly retweeted scholarly articles. Employing the altmetric.com dataset, we examine the tweets of highly retweeted research articles. We detected the obvious burst phenomenon in the social media diffusion of highly retweeted articles. The burst accumulates rapidly, but the recession is also fast. The duration of the burst is about 20 hours, and the diffusion scale is 25-1000. For the sample paper of this study, it occupies more than 60% of the total dissemination volume, even up to 100%. We found that medical articles are absolute hot spots on twitter and highly-tweeted papers are usually topics of interest to both the general public and scientific audience. The result of topological analysis shows that the retweet networks have a limited height but a wide range of widths which is a kind of very typical mass communication, rather than interpersonal communication.

Introduction

The social web has already emerged as a significant medium for the widespread distribution of news and messages. With hundreds of millions of users worldwide and their generated social network, user-generated content spreads quickly once released on a social web. Users of social media platforms like Twitter and Facebook receive, share, and disseminate content at a massive scale, which provides the possibility for content to accumulate tons of attention. However, the accumulation is not a one-time process, it follows a temporal trend (Budak, & Abbadi, 2011). Trends in social networks have been a major focus of interest among researchers. Leskovec, Backstrom & Kleinberg (2009) observe a typical lag of 2.5 hours between the peaks of attention to a phrase in the news media and in blogs respectively. For social media, Onnela & Reedtsochas (2010) analyze the factors affecting the diffusion trend of applications in Facebook, while Asur, et.al (2011) investigates the formation, persistence, and decay of trends on hot topics on Twitter. The online social interactions like retweet, reply, comment and mention would then boost the content propagation, spread different ideas and synchronize the collective attention of massive individuals, which might finally produce trends in online social media (Borner, Maru, & Goldstone, 2004; Crane, & Sornette, 2008; Wu, et.al, 2011; Zhang, Zhao, & Xu, 2016).

Existing trends research of social web focuses on information such as news, messages or other behavioral information, nevertheless, the popular trend of academic information like scholarly articles on social media has rarely been studied (Sugimoto, et.al, 2017; Giannetti,

2018). Some research observed the growth of visiting or download trends of scholarly articles as a result of social media division. (Allen, et.al, 2013; Wang, Fang, & Guo, 2016; Wang, Xu, & Fang, 2016). More research attempts to link the social media buzz of scholarly articles to the final citation trend or forecast, but there is no clear conclusion that this link inevitably exists. (Eysenbach, 2011; Wouters, & Costas, 2012; Costas, Zahedi, & Wouters, 2015; Thelwall, & Nevill, 2018). Different from these articles, our research will turn to the temporal diffusion trend of scholarly articles on social media. More specifically, in this paper, we are going to observe the time curve of social media popularity and study the burst phenomenon of growth.

Burst is a fairly common phenomenon in various statistics in real life. Most processes driven by individual human actions have burst phenomenon (Barabási, 2005; Barabási, & Gelman, 2010), which are reflected as the existence of one or more peaks on the time series curve of statistics (Cha, Mislove, & Gummadi, 2009). On social media like Twitter and Facebook, the dissemination of scholarly articles has obvious short-term characteristics. Newly published articles can quickly attract explosive attention, but the enthusiasm for such attention comes and goes quickly (Wang, Fang, & Guo, 2016). Retweeting as one of the manifestations of this attention may also have burst phenomenon. Few studies have been done to analyze this phenomenon specifically during the diffusion of scholarly articles, so we want to fill this gap. The research questions are as follows:

1. Whether highly retweeted scholarly articles have burst phenomenon in diffusion on the social web?
2. If there is a burst, what are the characteristics of these bursts?
3. What is the mechanism of burst: interpersonal communication or mass communication?

Data

In this paper, we use the time series of tweets to measure the diffusion of highly retweeted scholarly articles (The types of articles are limited to research papers and reviews). Thus, the tweets data of scholarly articles is employed as the dataset and we focus on the scholarly articles with more than 1000 times of retweets. The metadata counting 687,168 tweets comes from Altmetric.com including tweeting content, time stamp, tweeting account and other information. The time ranges from Jun 2011 to October 2017. After data cleaning (filtering error data and incomplete data), the final amount of data used for the study is 584,205 of 244 articles.

Results

Peak recognition

The burst means there is a surge in tweeting an article within days or even hours, which appears as one or more peaks on the temporal tweeting trend. The first step is to identify the peaks. Before that, we need to lay some groundwork to visualize and observe the tweeting trend and retweet network. For each article, we construct the retweet network based on the retweet relationship between Twitter accounts. Drawing tweeting trend by the time series of tweets data, and then observe whether there are peaks for the curve of temporal tweeting trend. If it exists, it means that there is burst diffusion in retweet network.

According to the observed results, we divide the tweeting curve into three types: unimodal curve, bimodal curve, and polymodal curve. As shown in Figure 1 to 3.

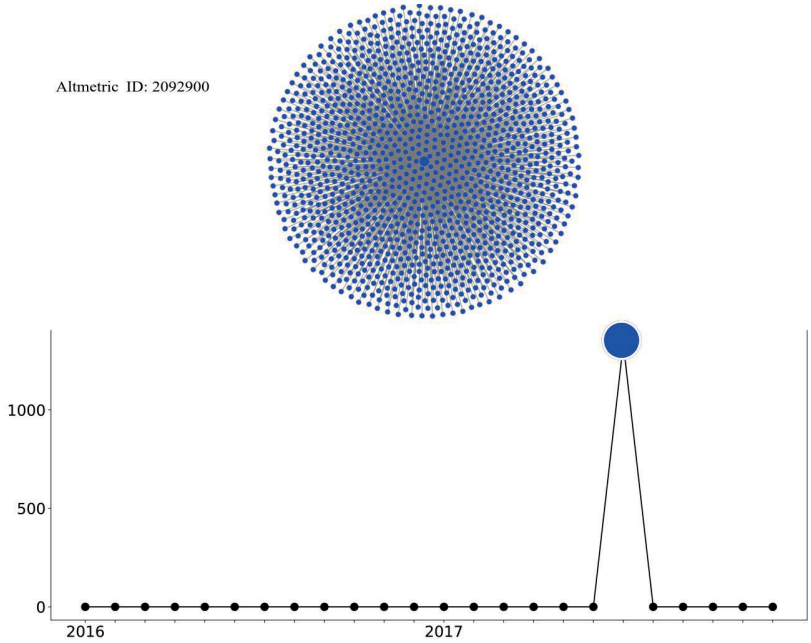


Figure 1. Sample of unimodal curve and the retweet network of the peak

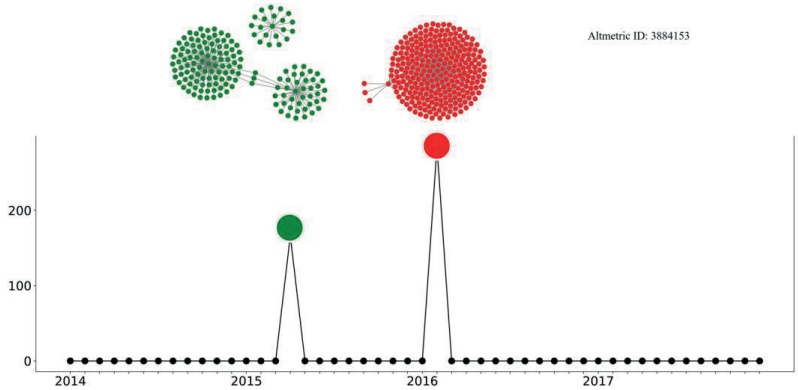


Figure 2. Sample of bimodal curve and the retweet network of the two peaks

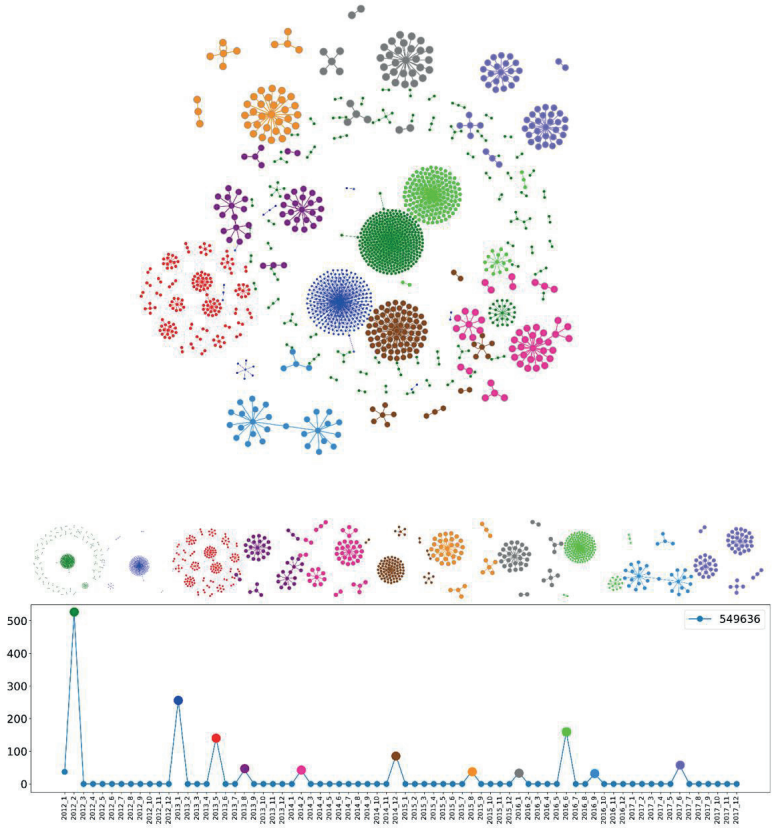


Figure 3. Sample of polymodal curve and the retweet network of the multiple peaks

Figure1 to 3 displays both the tweeting curve (temporal tweeting trend) and corresponding retweet network of one or more peaks from 3 sample paper (altmetric.com id:2092900, 3884153 and 549636). Each peak is marked with different color, and the retweet subnetwork represented by each peak is drawn above the curve. Articles with a unimodal curve have only once attracted large-scale attention but quickly returns to normal on social media during the reporting period. While bimodal or polymodal curve means that articles were mentioned again or received sustained attention over a period of time. Generally, there are one or more peaks scattered on the tweeting curve of an article. We design an algorithm to identify and extract these peaks. Firstly, calculating the mean of non-zero values in the daily tweeting time series and then finding all values greater than 2 times of the mean. Here we only care about those peaks with sufficient tweets, so we set the threshold as 2 times of the mean. Next, extracting the peak interval according to the following rules: assuming that

the tweeting sequence is A[N] (here N is time in days), find all ternary sequences that satisfy the value A[P] > A[P-1] and A[P] > A[P+1], then the tweets in the time period [P-1,P+1] is considered as a burst. Finally, we extracted a total of 988 peaks for the 244 articles.

Content analysis of burst characteristics.

What domain has attracted so much attention, and what are these articles about? To answer this question, the research areas and tweets of all 244 articles were analyzed. We browsed each article and divided the article into eight categories: medicine, biology, social sciences, computer science, physics and chemistry, psychology, multidisciplinary and others. The classification standard refers to the classification in the Scopus database. Figure 4 gives the proportion of these categories.

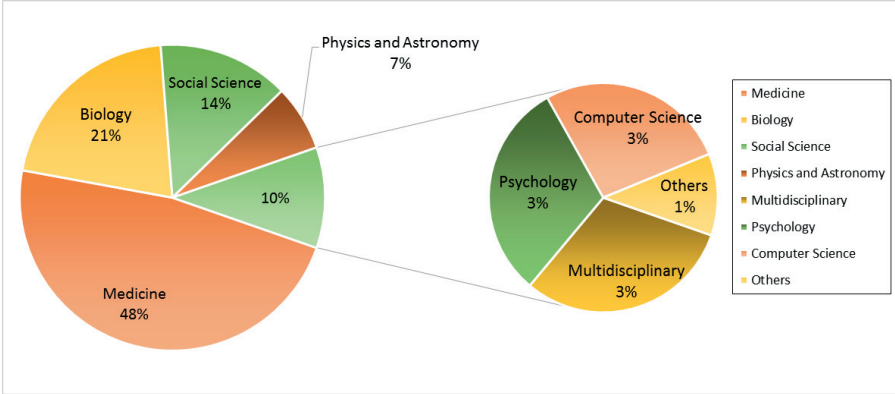


Figure 4. Research areas and its proportion

Using LDA (Latent Dirichlet allocation) topic model (Blei, Ng, & Jordan, 2003), we generated topics words from 584,205 tweets according to the categories. LDA is a generative statistical model that is widely used to identify hidden subject information in large document sets or corpora. It is an unsupervised model that only needs the text set and the number of topics. After training, the word distribution under each topic can be output. Here we train it to identify the topic information implied in the retweet text. Like Table 1 shows, we set 7 topics (from Medicine to Computer Science) and then selected the top 10 representative key words from the output word distribution.

Table 1. Topic words of retweets

Medicine	*Biology*	*Social Science*	*Physics and Astronomy*	*Multidisciplinary*	*Psychology*	*Computer Science*
Health	Drug	Coup	Quantum	Quantum	Phd	Neural
Risk	Life	CIA	Wave	Network	Mental	Game
Public	Tree	CNN	LIGO	Brain	Health	Network
Vaccine	Human	Gun	Qubit	System	Share	Deep
Zika	Infant	Science	Star	Circuit	Student	Search
Virus	CRISPR	Trump	Planet	Graph	Love	Photo
Cancer	Malaria	Law	Proxima	Volume	Mind	Machine

Clinic	Cell	Ethnic	Black	Energy	Press	Alphago
Sepsis	Embryo	Gender	Hole	Muscle	Control	Feature
Surgery	Ape	Reform	Orbit	Nano	Demand	Bigdata

Combined with Figure 4 and Table 1, we found that medical articles are absolute hot spots on twitter. It occupies almost half of all articles (48%). People are generally concerned about health, risks, viruses, cancer, etc. Biology ranks second, accounting for 21% and its topics are drugs, life, infant, CRISPR, cells, etc. Following is the social sciences accounted for 14% whose topics are related to issues such as politics, law, and racial discrimination. Physics and astronomy research also accounted for 7%, of which quantum, gravitational waves, black holes, etc. are hot topics of discussion. The total number of articles in the above four areas accounted for 90% of the total. It seems that highly-tweeted papers are usually topics of interest to both the general public and scientific audience.

Indicator analysis of burst characteristics.

In this section, we apply four indicators to measure the burst of the temporal tweeting curve, which are diffusion size, bursts proportion, rise time and decay time.

· Diffusion size. It describes the scale of propagation during a burst, that is, how many people participated in tweeting totally.

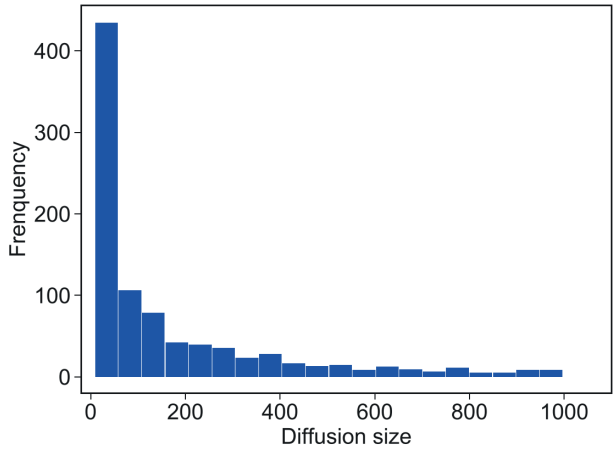


Figure 5. Distribution of diffusion size

Figure 5 shows the distribution of diffusion size of all bursts. The size of diffusion varies from more than 10 to nearly 1000. This may be related to the influence of the publishers or retweeting accounts. A high-impact account tends to have a large number of fans, which makes it easy to get a lot of retweets.The diffusion size of polymodal bursts is concentrated in the range of 0 to 200, while that of unimodal or bimodal bursts forms the long tail in this

figure.

- Bursts proportion. It measures the proportions of all bursts in the dissemination of an article. Generally, there are one or more peaks on a tweeting curve, which means that for the entire propagation process of an article, one or more bursts usually occur. We are concerned about the contribution of the scale brought about by these bursts. The calculation method is the ratio of the all bursts' diffusion size in an article to the total diffusion size of this article.

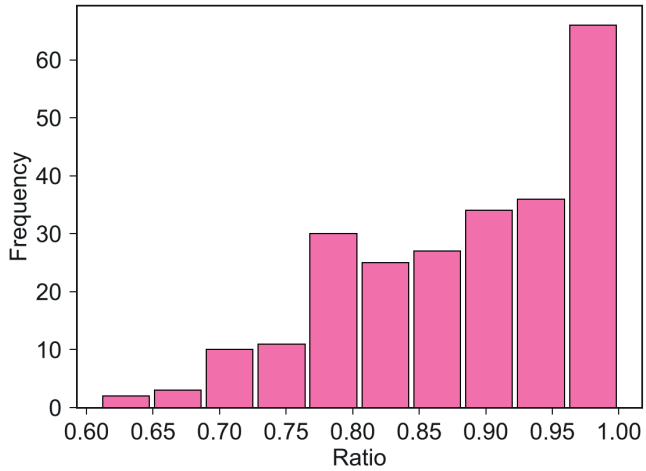


Figure 6. Bursts proportion distribution of all articles

Here we calculate the bursts proportion of all 244 articles (Figure 6). Nearly 90% of the articles have more than 75% of the bursts proportion. The smallest percentage is higher than 60%, while highest can be 100%. That is to say, the bursts are the main part of the dissemination of these articles, which contributed 60% - 100% of the total dissemination scale. This confirms the conclusion of the preliminary study that attention on social networks is not slowly accumulated, but suddenly erupts.

- Rise time. It measures the time from the start to the peak of the burst.

- Decay time. The decay time is the time from the peak time to the time when the number of retweeting people reaches 75% of the diffusion size. It helps to understand the duration of the burst.

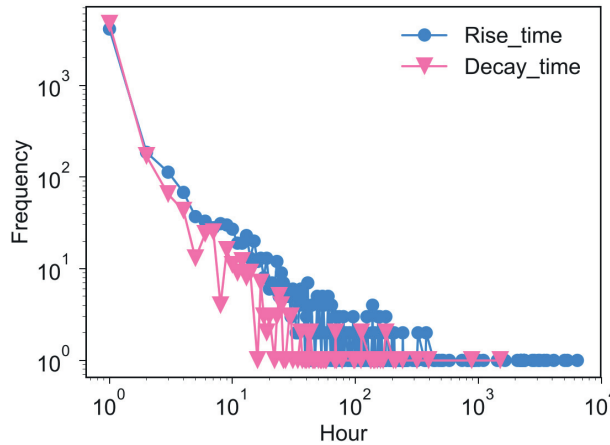


Figure 7. Distribution of rise time and decay time in hours

Figure 7 shows the distribution of rise time and decay time for all bursts in hours. As shown in the figure, both the rise and decay distribution curves are long tailed, which means that both the rise time and the decay time are generally less than 10 hours. Most bursts do not last very long, reaching their maximum value in only a few hours, and then decaying rapidly and faster in the next few hours.

How and why the burst occurs: topological analysis of retweet network.

The topology statistics of the retweet network can help understand how and why the burst occurs. By tracing the retweeting relationship, we can get the propagation paths, and then establish the retweet network. We use the hierarchical tree model to characterize propagation paths (Hanneman, & Riddle, 2005), as Figure 8 shows, the information published by the root node O is sequentially spread to a total of 10 leaf nodes (a, b, c, d, e, f, g, h, i, j). An edge from node A to node B is added to the tree only when B retweets content from A. In general, the propagation path is a simple tree structure. But in real data, because one person can retweet information from two people at the same time, the propagation path is not a tree structure. However, a complicated propagation path can be decomposed into several simple propagation trees. In this case, the propagation pattern of a single tweet will contain multiple trees and form a forest. (Rodrigues, et.al, 2011). Fortunately, it doesn't affect what we're going to do next.

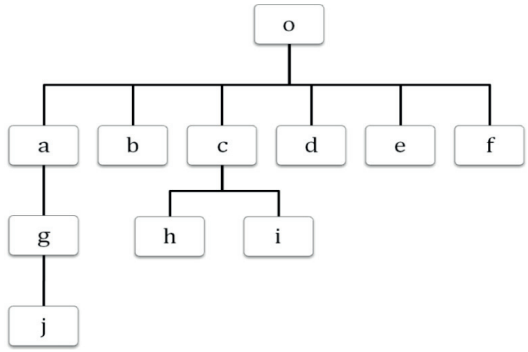


Figure 8. Hierarchical Tree Mode

There are 3 topological statistics, which are height, width and propagation motif.

· Height of retweet network. It is defined as the maximum number of hops from root node to a leaf node in all propagation trees. For example, the height of retweet network in Figure 8 is 3.

· Width of retweet network. It is defined as the maximum number of nodes at each leaf layer level in all propagation trees. For example, the number of the leaf layer in the network of Figure 8 is 3, and the first layer has up to 6 nodes, so the width of the network is 6.

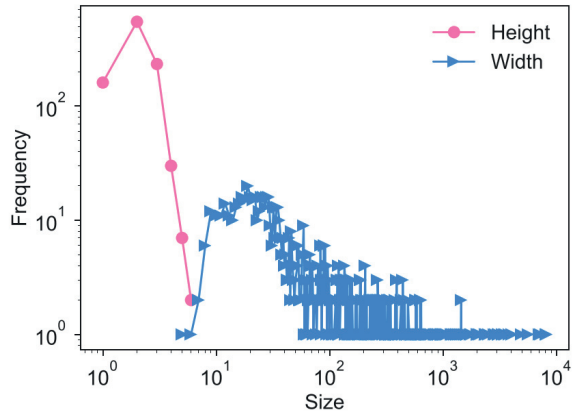


Figure 9. Distribution of height and width

We calculated the height and width of the retweet network generated by all bursts (Figure 9). The distribution of height and width indicate the propagation pattern of the burst. The propagation height of most bursts does not exceed 3, and the maximum propagation height does not exceed 10, while the propagation width is widely distributed from 10 to 1000. This means that the burst of dissemination of scholarly articles may be mainly based on mass communication. To validate this hypothesis, we use propagation motif to analyze the retweet network structure.

· Propagation motif: A motif refers to a structural model. Here we refer to a basic tree structure in the hierarchical tree model. Different motifs represent different modes of communication. In order to facilitate program operation, we use the following two motifs to denote interpersonal communication and mass communication. Like Figure 10, Motif I denotes mass communication: information is spread from one node to multiple nodes simultaneously. On the contrary, Motif II means that information is transmitted sequentially between nodes along a straight line. We call this model interpersonal communication. By calculating the number of these two structural motifs in all retweet networks, we can clarify the propagation mode of the bursts.

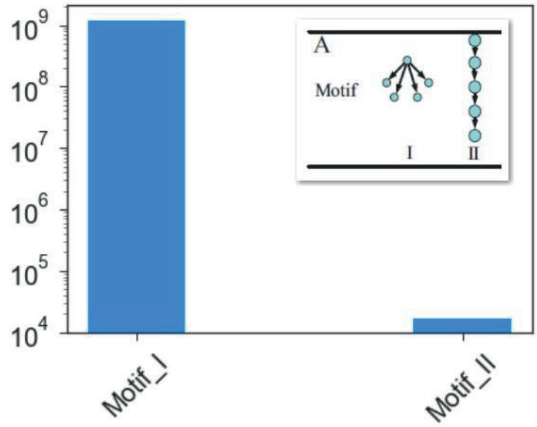


Figure 10. Two motifs of bursts

According to Figure 10, the number of Motif I is up to 5 orders of magnitude higher than Motif II. Therefore, it can be considered that the mode of communication here is mainly mass communication.

Discussion

In this study, the initial three research questions have been preliminarily answered. The diffusion of highly retweeted papers on twitter is not slowly accumulated, but there is a massive amount of bursts. Although these bursts have a short duration, only about 20 hours of rise and decay every time, they still have a diffusion scale of 25-1000. More importantly, our research found that the bursts brought most of the diffusion, usually above 75%, some even reach 100%. This illustrates that the burst plays a vital component in the dissemination of scholarly articles, and its size and quantity have a decisive impact on the overall diffusion scale. The topological analysis of the retweet network helps to explain the cause of the burst. In this study, we assume that the burst is affected by two main factors: the speed of information transmission and the length of the propagation path. The speed of information transmission on the Internet is so fast that it does not make a significant difference in the dissemination effect. Therefore, only the propagation path length really works. A burst shown as getting a large amount of attention in a short period of time means that there are a large number of short information propagation paths in the retweet network.

Our results support this assumption: the height of retweet network is very small but the width can be very large, and there is a great deal of Motif I with short propagation paths. The diffusion way of the burst is typical mass communication, which is pushed by one or more high-impact accounts transmit to a large number of fans synchronously. It shortens the time of dissemination and makes it possible to obtain a large amount of retweet rapidly. This kind of retweet heat brought by the fans can only be maintained for a short period of time. Although it comes very quickly, it then decays faster, which leads to a significant peak on the retweet curve: a burst happened.

References

Budak, D., & El Abbadi, A. (2011). Information diffusion in social networks: Observing and influencing societal interests. PVLDB, 4(12), 1-5.

Allen, H. G., Stanton, T. R., Pietro, F. D., & Moseley, G. L. (2013). Social Media Release Increases Dissemination of Original Articles in the Clinical Pain Sciences. PLOS ONE, 8(7).

Asur, S., Huberman, B. A., Szabo, G., & Wang, C. (2011). Trends in Social Media: Persistence and Decay. International conference on weblogs and social media.

Barabási, A. (2005). The origin of bursts and heavy tails in human dynamics. Nature, 435(7039), 207-211.

Barabási, A. L., & Gelman, A. (2010). Bursts: The hidden pattern behind everything we do. Physics Today, 63(5), 46.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. Journal of Machine Learning Research, 993-1022.

Borner, K., Maru, J. T., & Goldstone, R. L. (2004). The simultaneous evolution of author and paper networks. Proceedings of the National Academy of Sciences of the United States of America, 5266-5273.

Cha, M., Mislove, A., & Gummadi, K. P. (2009). A measurement-driven analysis of information propagation in the flickr social network. The web conference, 721-730.

Crane, R., & Sornette, D. (2008). Robust dynamic classes revealed by measuring the response function of a social system. Proceedings of the National Academy of Sciences of the United States of America, 105(41), 15649-15653.

Eysenbach, G. (2011). Can Tweets Predict Citations? Metrics of Social Impact Based on Twitter and Correlation with Traditional Metrics of Scientific Impact. Journal of Medical Internet Research, 13(4).

Giannetti, F. (2018). A twitter case study for assessing digital sound. Journal of the Association for Information Science and Technology, 69(5), 687-699.

Hanneman, R. A., & Riddle, M. (2005). Introduction to social network methods.

Leskovec, J., Backstrom, L., & Kleinberg, J. M. (2009). Meme-tracking and the dynamics of the news cycle. Knowledge discovery and data mining.

Onnela, J., & Reedtsochas, F. (2010). Spontaneous emergence of social influence in online systems. Proceedings of the National Academy of Sciences of the United States of America, 107(43), 18375-18380.

Costas, R., Zahedi, Z., & Wouters, P. (2015). Do “altmetrics” correlate with citations? Extensive comparison of altmetric indicators with citations from a multidisciplinary perspective. Journal of the Association for Information Science and Technology, 66(10), 2003-2019.

Rodrigues, T., Benevenuto, F., Cha, M., Gummadi, P. K., & Almeida, V. A. (2011). On word-of-mouth based discovery of the web. Internet measurement conference.

Sugimoto, C. R., Work, S., Lariviere, V., & Haustein, S. (2017). Scholarly use of social media and altmetrics: a review of the literature. Journal of the Association for Information Science and Technology, 68(9), 2037-2062.

Thelwall, M., & Nevill, T. (2018). Could scientists use Altmetric.com scores to predict longer term

citation counts. Journal of Informetrics, 12(1), 237-248.
Wang, X., Fang, Z., & Guo, X. (2016). Tracking the digital footprints to scholarly articles from social media. Scientometrics, 109(2), 1365-1376.
Wang, X., Xu, S., & Fang, Z. (2016). Tracing Digital Footprints to Academic Articles: An Investigation of PeerJ Publication Referral Data. arXiv: Digital Libraries,.
Wouters, P., & Costas, R. (2012). Users, narcissism and control: tracking the impact of scholarly publications in the 21st century (pp. 847-857). Utrecht: SURFfoundation.
Wu, S., Tan, C., Kleinberg, J. M., & Macy, M. W. (2011). Does Bad News Go Away Faster. International conference on weblogs and social media.
Zhang, L., Zhao, J., & Xu, K. (2016). Who creates Trends in Online Social Media: The Crowd or Opinion Leaders? Journal of Computer-Mediated Communication, 21(1), 1-16.

Scholarly book publishers and their promotional activity on Twitter

Yajie Wang[1,2] and Alesia Zuccala[2]

[1,2] *yajie.wang@hum.ku.dk*, [2]*a.zuccala@hum.ku.dk,*
[1]*WISE Lab, Institute of Science of Science and S&T Management, Dalian University of Technology, Dalian, 116024, (China)*
[2]*Dept. of Information Studies, University of Copenhagen, Njalsgade 76, 2300 København S, (Denmark)*

Abstract

This is a research-in-progress paper concerning the activity of book publishers and authors on Twitter. The work is based on a sample of book titles (*N*=2,672) published between 2014 and 2018 (up to and including October 1, 2018), and extracted from the Web of Science Book Citation Index. Our motivation is to learn more about book publishers as promotional agents. While we know that scholars often use Twitter to promote their research articles, and may do so with monographs/series titles, little is known about publishers. Scholars may Tweet about their monographs to gain visibility and increase citation counts. Publishers of the same titles might also do the same, but they are co-stakeholders in this communication process with other interests. With our pilot dataset, we have found that only 42% of the titles retrieved from the BKCI possessed DOI's and had received Twitter mentions. We have also found that if the author of a book initiates a Twitter mention, other general Twitter users are likely to make follow-up/re-tweets about the title. Book publishers, on the other hand, do often initiate Tweets, but they are less often as likely to make a re-tweet of a book title if the author or another user initiates.

Introduction

Twitter is one of few social media platforms to attract a variety of academic users. According Haustein (2018) "academic careers are no longer shaped only by peer-reviewed papers, citation impact and impact factors" because "university managers and funders now also want to know how researchers perform on social media and how much their work has impacted society at large". A growing body of literature has thus been produced and reviewed concerning social media (Sugimoto et al., 2017), with a strong focus on Tweets made by scholars, or communities of scholars, and their research articles (e.g., Bowman, 2015; Darling et al., 2013l; Didegah et al., 2016; Haustein et al., 2014; Haustein et al., 2015; Robinson-Garcia et al., 2017).

The focus of this research is on a specific document type - i.e., scholarly books, including both book title authors and publishers as Twitter users. Our motivation for examining book publishers is based on the fact that they are as much a part of the scholarly communication continuum as scholars, albeit with a different stakeholder position in terms of promotional activity. While it helps the reputation of a scholar if his/her books are visible and achieve a broader impact via social media, it can also do the same for the publishers who print the books. Publishers have a simple mandate and that is to market their 'products' and thrive both commercially and economically.

We are thus interested in understanding a) how scholarly publishers use Twitter to announce and promote their book titles, b) what they say about the book titles, and c) how this relates to tweets made by the same book title authors (or editors), and general uptake of the title overall on Twitter by 'everyone else'. We also want to know if the initial 'marketing' activity on Twitter by a publisher relates to some degree with the book's scholarly uptake or 'citedness'. Overall, the aim is to learn more about the scholarly book publisher as promotional agent and its role in relation to a book's visibility and impact.

Related Literature

Research concerning book publishers on Twitter has focused almost exclusively on the 'trade' publishing industry, and not on scholarly publishers or university presses. While the

two publisher types differ in terms of customer base (i.e., the latter printing books for academic libraries and scholars), trade publishers do also publish educational material for schools, colleges and/or universities (e.g., The Balance Careers, 2019). Still, much of the 'hype' surrounding Twitter's use for marketing purposes has been targeted towards trade book authors and trade publishers (Abbot, 2009 March 30). According to Thoring (2011) Twitter "introduces an unparalleled opportunity [for trade publishers] to interact with existing and potential customers" even though the "level of noise is incredibly high and constitutes a challenge to marketers (p. 144). Following an analysis of 194 UK-based trade publishers (5 large, 18 medium-sized, 171 small), Thoring found that tweets to books were made "to inform followers about previews, pre-orders, book trailers and incidents relating to a book or to recommend titles" yet in terms of where the Tweets led, she concluded that "most ...contained hyperlinks, but there was no dominant type of destination" (p. 154).

So far, there is little evidence to suggest that *scholarly publishers/university presses* are less interested than their trade counterparts in using social media. However, most studies surrounding scholarly books have focused on 'mentions' to individual book titles and not on the publishers producing these titles (e.g., Hammarfelt, 2014; Torres-Salinas et al., 2017; Torres-Salinas et al., 2018). Prior to the development of data providers like Altmetrics.com and PlumX, a search for full or partial book titles had to be carried out manually via individual platforms (Hammarfelt, 2014). Now, Altmetric.com allows researchers to find mentions to all kinds of research outputs, including books, book chapters, and even datasets. Torres-Salinas et al. (2018) investigated these outputs on Altmetric.com, and found that only 5 percent were to books, followed by 2.3 percent to book chapters. The authors also point to the fact that Altmetric.com traces mentions via Digital Object Identifiers (DOI), and that these are strongly associated with articles, and less frequently applied to monographs/books. A similar, if not complementary problem is that few DOI's appear for books on commercial citation indexes (i.e., Scopus and Web of Science) (Gorraiz et al., 2016).

Barring this DOI problem, research opportunities pertaining to books are opening up in general. The ISBN, for instance, functions as another form of identifier. Despite the difficulties associated with recognizing books with several ISBNs, it has been used for a few book-oriented studies (Kousha et al., 2017; Zuccala & Cornacchia, 2016). It is therefore possible to trace social media mentions to a book using its title and/or ISBN, or even a hyperlink to its publisher webpage, if a DOI is not available (Torres-Salinas et al., 2018).

In comparison to journal articles, books represent a lower percentage of what is indexed on Altmetric.com, yet Konkiel and Adie (2018) state that "as of July 2018, [this platform] has tracked attention for more than 829,000 books and 80,000 book chapters across a wide range of subjects" (p. 2). Similar to journal articles, statistics show that "more than 70% of [mentions to books] occur on Twitter" (p. 3). And, with all of the publishers and presses, (i.e., 4,055) present, it is not surprising that Oxford University Press and Routledge are the highly featured. Earlier research pertaining to university press rankings shows that these are amongst some of the most prominent and powerful publishing houses in certain humanities fields, like history (Zuccala et al., 2014).

Given the type of research that has taken place concerning scholarly book publishers and social media, we see a strong potential now for focusing on the role of publishers via the Twitter communication cycle. Again, scholars may 'promote' themselves and their work at any time via Twitter or other social media, but publishers have their own opportunities to consider, and represent another form of trust for readership communities online.

Methodology

Data collection: A selection of monographs (B), and book series (S) (N=2,672) with the publication years (PY) 2014 to 2018 were extracted from the Web of Science Book

Citation Index (Clarivate Analytics) on October 1st, 2018. Table 1 shows the skewed frequency distribution of citations received by the titles in our dataset.

Table 1: frequency distribution of the book titles

	≥ 20 citations	10 to 20 citations	10 to 5 citations	5 to 1 citations	0 citations
2,672 Titles (Monographs and Book Series Titles)	65	124	243	707	1533

In order to determine the Twitter activity associated with these books, we utilized Altmetric Explorer at Altmetric.com. With this platform it is possible to search for 25,000 different scholarly identifiers, including DOIs, ISBNs, PubMed IDs, arXiv IDs, etc. However, for our sample of books we focused primarily on titles possessing a DOI or ISBN, and collected Twitter statistics utilizing these identifiers. As noted earlier, many books included in commercial indices like Web of Science or Scopus, do not possess a DOI (Gorraiz et al., 2016); thus, we were not surprised to find with our dataset that this identifier was present in only 42% of the cases. ISBNs, on the other hand, were 100% present.

From the pilot sample of 2,672 book titles, we found that a total of 1,098 had received mentions on Twitter. We then manually checked to see what type of account had made the tweets, specifically, if it was the author of the book or if it was the publisher, or another account. Note that some publishing industry giants often have multiple Twitter accounts at the same time in order to facilitate news and/or achieve a stronger promotional effect. These multiple accounts may be set up to belong to different disciplines or fields, whereby the suffix of the Twitter account often adds this information. Table 2 presents one example below, for the international academic and trade publisher, Palgrave Macmillan.

Table 2. Palgrave Macmillan's Twitter accounts.

General Twitter Account	**Business, Economics and Finance**	**Twitter - Humanities**	**Twitter - Social Sciences**
@Palgrave	@PalgraveBiz	@PalgraveCultMed	@PalgraveCrim
@PalgraveJournal	@PalgraveEcon	@PalgravePhil	@PalgraveEducate
@PalgraveLibrary	@PalgraveFinance	@PalgraveHistory	@palmacpolitics
		@Palgrave_Ling	@PalgravePsych
		@PalgraveLit	@PalgraveSoc
		@PalgraveTheatre	@PalgraveGeoEnvi

For publishers similar to the one shown in Table 2, we performed additional checks by directly entering the publisher's name in the Twitter search field. Then from our search results, we examined the Twitter account's introduction, the amount of attention it was receiving, and whether or not the account was attributed also to the publisher's official homepage. The drawback to this type of manual test is that it depends on the judgement of the researcher; however, it was necessary for ensuring the accuracy of our data.

After obtaining the Twitter account for each publisher, we then explored the publisher tweets in relation to the 1,098 titles. In doing so, we focused on Tweet frequencies, including tweet initiation versus re-tweeting, and information about whether the book's title, DOI, ISBN

or a website was included in the mention. If the originator of the first tweet was the publisher of the book title, we collected further data pertaining to the overall Twitter profile, such as overall tweets, number of followers, tweets per day, percentage of re-tweets, replies to the tweets, average number of links, hashtags, and the proportion of tweets that had been favoured or re-tweeted by others. Overall, there was a judgement process applied to the data collection, which is outlined in Figure 1.

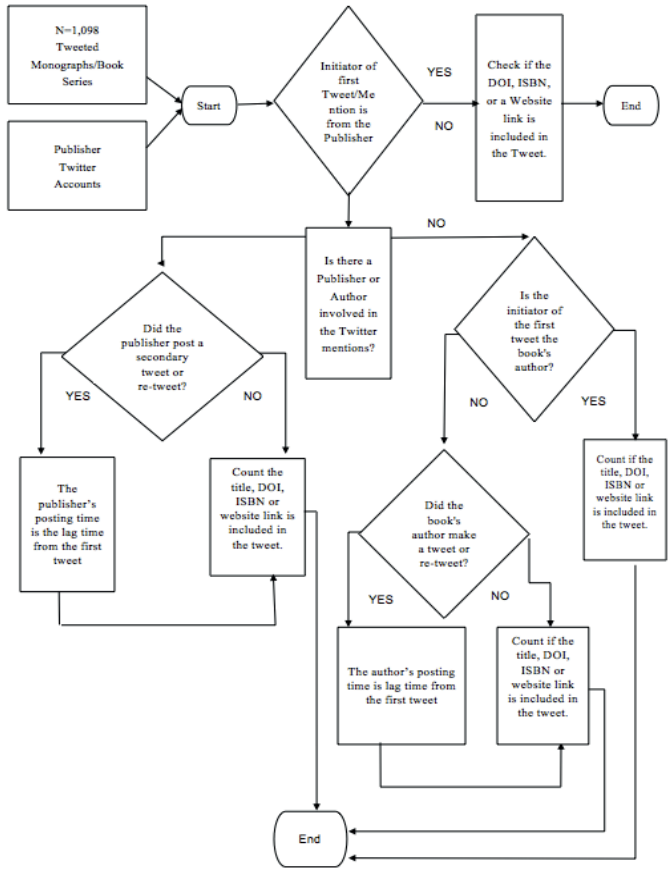


Figure 1: Judgment process applied to Twitter data collection.

Results and Preliminary Conclusions

Tables 3 and 4, as well as Figure 2, below, present our preliminary findings concerning tweets given to the 1,098 book titles (monographs and book series published in 2014 to 2018). Note that although the book authors and publishers are involved in producing Twitter mentions to their titles, other general Twitter users tend to mention the titles more frequently. Our pilot study also showed that when we compare the average citation rates to books, there does not seem to be any hint of a promotional advantage. By this, we mean that it is too early to suggest that if a publisher mentions a book (via title, ISBN or DOI) on Twitter that it relates to a higher citation rate. Also, because of the time it takes to cite a book

versus the time it takes to Tweet, we are aware of the limitations attached to comparing the two indicators (e.g., Thelwall et al., 2013).

What we do see; however, is that initiator Tweets made by authors are more likely to increase further mentions on Twitter or generate re-tweets by other Twitter account users. While publisher accounts were found to make a lot of initial tweets, very few make a re-tweet concerning their own books. Further research is required to investigate more fully the role that scholarly book publishers play on Twitter.

Table 3. First tweets to the book titles and type of Twitter account.

Account making the first mention or "Initiator" of the Tweet	Number of book titles	Average citation rate of the titles	Average number of mentions from all users
Publisher	138	2.01	3.26
Book Author (s)	129	1.42	9.80
Other Twitter user (e.g., members of the public, other scholars/scientists/ practitioners)	831	3.14	3.33

Table 4. Initiator tweets, or first mentions to the book titles in terms of percentages.

Type of Tweet	Type of Twitter Account		
	Publisher	Author	Other Twitter Users
Original tweet	92%	79%	76%
Re-tweet	8%	21%	24%

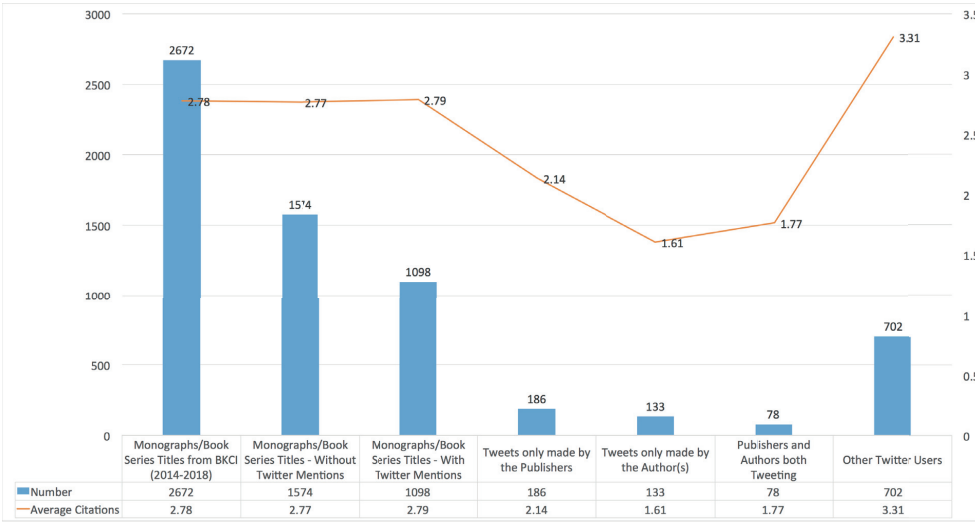


Figure 2. Trends related to Tweets to Monograph/Book Titles published 2014-2018.

References

Abbott, C. (2009 March 30). To Tweet or not to Tweet: twitter basics for publishers and authors: promoting books in 140 characters or less. *Publishers Weekly, 256(13)*. Retrieved November 26, 2018 from http://www.publishersweekly.com/pw/print/20090330/2722-to-tweet-or-not-to-tweet-twitter-basics-for-publishers-and-authors-.html.

Bowman, T. D. (2015). Differences in personal and professional tweets of scholars. *Aslib Journal of Information Management*, *67*(3), 356–371. https://doi.org/10.1108/AJIM-12-2014-0180

Darling, E. S., Shiffman, D., Côté, I., & Drew, J. A. (2013). The role of Twitter in the life cycle of a scientific publication. *Ideas in Ecology and Evolution*, *6*, 32–43. https://doi.org/10.7287/peerj.preprints.16v1

Didegah, F., Bowman, T. D., Bowman, S., & Hartley, J. (2016). Comparing the characteristics of highly cited titles and highly alted titles. In *21st International Conference on Science and Technology Indicators* (pp. 1190–1195). Valencia, Spain. Retrieved from http://ocs.editorial.upv.es/index.php/STI2016/STI2016/paper/viewFile/4543/2327

Gorraiz, J. Melero-Fuentes, D., Gumpenberger, C. & Valerrama-Zurián, J. (2016). Availability of digital object identifiers (DOIs) in Web of Science and Scopus. *Journal of Informetrics, 10(1)*, 98-109.

Hammarfelt, B. (2014). Using altmetrics for assessing research impact in the humanities. Scientometrics, 101(2), 1419-1430. https://doi.org/10.1007/s11192-014-1261-3.

Haustein, S., Bowman, T. D., Holmberg, K., Peters, I., & Larivière, V. (2014). Astrophysicists on Twitter: An in-depth analysis of tweeting and scientific publication behavior. *Aslib Journal of Information Management*, *66*(3), 279–296. https://doi.org/10.1108/AJIM-09-2013-0081

Haustein, S., Costas, R., & Larivière, V. (2015). Characterizing social media metrics of scholarly papers: The effect of document properties and collaboration patterns. *PLoS ONE*, *10*(3), e0120495. https://doi.org/10.1371/journal.pone.0120495

Kousha, K., Thelwall, M., and Abdoli, S. (2017). Goodreads reviews to assess the wider impacts of books. *Journal of the Association for Information Science and Technology*, *68(8),* 2004-2016.

Robinson-Garcia, N., Costas, R., Isett, K., Melkers, J. and Hicks, D. (2017). The unbearable emptiness of tweeting – about journal articles. *PLOS One, 12(8),* 1-19. https://doi.org/10.1371/journal.pone.0183551.

Sugimoto, C. R., Work, S., Larivière, V., & Haustein, S. (2017). Scholarly use of social media and altmetrics: a review of the literature. *Journal of the Association for Information Science and Technology, 68*(9), 2037–2062. https://doi.org/10.1002/asi.23833

The Balance Careers (2019). The big 5 trade book publishers. Retrieved November 29 2019 from https://www.thebalancecareers.com/the-big-five-trade-book-publishers-2800047.

Thelwall, M., Haustein, S., Lariviere, V., & Sugimoto, C. R. (2013). Do altmetrics work? Twitter and ten other social web services. PLoS ONE, 8(5), doi:10.1371/journal.pone.0064841.

Thoring, A. (2011). Corporate Tweeting: Analysing the use of Twitter as a Marketing Tool by UK Trade Publishers. *Publishing Research Quarterly, 27(2),* 141-158. https://doi.org/10.1007/s12109-011-9214-7.

Torres-Salinas, D., Gumpenberger, C. and Gorraiz, J. (2017). PlumX as a potential tool to assess the macroscopic multidimensional impact of books. *Frontiers in Research Metrics and Analytics, 2(5)*, 1-11. https://doi.org/10.3389/frma.2017.00005.

Torres-Salinas, D., Gorraiz, J. & Robinson-Garcia, N. (2018). The insoluble problems of books: what does Altmetric.com have to offer?" *Aslib Journal of Information Management, 70(6)*, 691-707.

Zuccala, A & Cornacchia, R. (2016). Data matching, integration, and interoperability for a metric assessment of monographs. *Scientometrics, 108(1),* 465-484.

Zuccala, A., Guns, R., Cornacchia, R., & Bod, R. (2014). Can we rank scholarly book publishers? A bibliometric experiment with the field of history. *Journal of the Association for Information Science and Technology, 66(7)*, 1333-1347.

Monetization Strategies of University Patents Through PAEs: an Analysis of US Patent Transfers

Stefania Fusco[1], Francesco Lissoni[2], Catalina Martinez[3], and Valerio Sterzi[4]

[1] sfusco@nd.edu
Notre Dame Law School, Notre Dame (United States)

[2] francesco.lissoni@u-bordeaux.fr
GREThA – UMR CNRS 5113, Université de Bordeaux, Bordeaux (France) &
ICRIOS-Università « L.Bocconi », Milan (Italy)

[3] catalina.martinez@csic.es
Institute of Public Goods and Policies, CSIC, Madrid (Spain)

[4] valerio.sterzi@u-bordeaux.fr
GREThA – UMR CNRS 5113, Université de Bordeaux, Bordeaux (France)

Abstract

The pressure to extract rents from academic research results has led many universities to file more patents and to rely on a growing range of monetization strategies including selling patents to Patent Assertion Entities (PAEs). We build a database of university patents granted by the USPTO and, for each of them, we collect information about the change of ownership. A first analysis of these data shows that about 12% of university patents have been transferred at least once (including reassignments to universities, hospitals, public research centres and governmental institutions) and only a minor part has been acquired by PAEs (the 0.3% of university patents). However, we also find that most transfers of university patents to PAEs occurred in the last ten years (3.4% of transfers). These acquisitions are largely concentrated in two large PAEs that acquired about 70% of all PAEs-acquired university patents: *Intellectual Ventures* and *Intellectual Discovery*. An econometric analysis on the characteristics of university patents transferred to PAEs shows that patents transferred to PAEs are of high quality, suggesting that PAEs cherry pick good patents for monetization purposes. However, PAEs acquire from universities older patents than those transferred to producing companies. This fact suggests that these transfers are not linked to technology transfer.

Introduction

Since the passage of the Bayh-Dole Act in 1980, which assigns the IP of patentable results from government-funded research to universities, the number of university patents has increased significantly in the United States (Merrill & Mazza, 2011). Moreover, in recent years, similar legislative acts have been adopted in other countries to strengthen institutional ownership controls and promote commercialization and technology transfer. Consequently, the number of university patents has increased in other countries as well.

Universities are also progressively more active in the patent market. Recent figures from USPTO show that universities are among the entities with the highest volume of outbound patent assignment transactions. Four universities (University of Pennsylvania, University of Alabama, University of Michigan and University of Colorado) are in the top five of patent assignors by number of transactions in the first two months of 2019 (IAM, 2019).

In addition to the increase in university patenting activity and the participation of universities in the patent market, there has also been a significant increase in the monetization activity of the relative patents. Universities have been engaged in technology transfer for decades, but studies seem to indicate that they have recently become more aggressive in trying to monetize

their patents through enforcement actions, licensing and transfers of their patents to other entities.

Specifically, some concerns have arisen about the possibility of using auctions and patent intermediaries such as Patent Assertion Entities (PAEs) for monetizing university patents.

Auctions are becoming more and more a way of commercialization chosen by universities. For example, university patents represent 20% of the business for Ocean Tomo, an IP merchant bank that organizes patent auctions (Ledford, 2013). Moreover, in 2014, Penn State University launched the first online auction of patent rights resulting from university research in the United States, offering about 70 patents to the highest bidder (Cahoy et al. 2016).

Transfers of university patents to PAEs, that is, companies whose exclusive business activity is to monetize patents through sales, licensing and litigation, have been also under scrutiny (Ewing & Feldman, 2012). PAEs have been accused by some observers to make new technology more expensive and using the patent system in a way that is contrary to the purpose for its creation. Contrary to universities and practicing entities (such as manufacturing companies), a PAE generally does not engage in research activity nor does it produce the goods or services covered by the intellectual property it controls. Several examples of transfers to PAEs have been publicly debated. One of the most notorious cases is the 2008 exclusive license of 50 patents of Caltech to a subsidiary of Intellectual Ventures (Ledford, 2013).

There could be several reasons for the increase in the university monetization activity mentioned above. In the U.S. context, Firpo and Meriles (2018) point out the fact that there have been reductions in the research funding provided to universities by the U.S. government; thus, universities need to finance research activities through other sources of income including the monetization of their patents. Moreover, they emphasize that university technology transfer offices are often not self-sufficient and that the enforcement of patents held by their universities helps them to increase revenues. Finally, based on Rai and Eisenberg (2003), a shifting of norms in the academia may have occurred, with universities becoming more inclined to hire professors who focus on applied research and whose results can be patented, as opposed to those engaging in basic research. The result of such hiring practices is a higher number of patents held by universities that ultimately leads to a higher level of monetization activity and, thus, to the development of patent thickets or an anticommons in relative fields (Firpo & Meriles 2018).

Since universities play an important role in both producing and disseminating knowledge, the way in which they monetize their patents might have a significant impact on society. Specific concerns have emerged in relation to access to university-created inventions that in many cases are funded by the public (Drivas et al., 2017; Thompson et al., 2018). Thus, it is important to study this phenomenon thoroughly. This research addresses this issue by providing a full analysis of the newly available data on transfers of university patents granted by the United States Patent and Trademark Office (USPTO) between 1990 and 2017 (Graham et al., 2018).

How often universities do transfer their patents to PAEs? Which are the characteristics of these patents? Our research is the first to address these questions in a detailed and systematic way. It covers the transfer of university patents granted by the USPTO over 28 years and reveals that such transfers are a growing phenomenon and do not seem to be associated to an increasing technology transfer to the private sector.

Data construction

We exploit information from the US Patent Assignment Dataset (PAD, Version 2017), which allows for identifying patent transfers registered at the USPTO. We then match the relevant patent data to those contained in the PatentViews database (www.patentsview.org) to retrieve important characteristics of the patents subject to a transfer, such as technological classes, citations, and age.

The exploitation of PAD data presents three main challenges. First, the patent applicants' names are not harmonized. Second, often the sector of the patent holder (private business enterprises, universities / higher education institutions, governmental agencies, individuals, etc.) is not reported. Third, changes in patent ownership may be the result of events that are not patent trades (mergers and acquisitions, collaterals, etc).
To address the first issue, we develop an algorithm to clean and consolidate patent applicants' names (disambiguation). Regarding the allocation of assignees by type of sector, we exploit the information from EEE-PAT to define the categories/sectors relevant for our subjects. Then, we list the identified assignees by sector and create an extensive list of PAEs in the private company category. Finally, patent applicant names from PAD are assigned to specific sectors when a similarity between the applicant name and a name on our list from EEE-PAT is found. These steps take care of the second challenge too.
To address the third issue, we select only records that PAD considers as new assignments; this leads to the exclusion of transfers of ownership deriving from mergers and acquisitions and transactions in which patents are used as collateral. For the purpose of our research, a patent is considered as "transferred" if and only if two unique disambiguated applicants have been registered consecutively on two different dates. We further clean the identified transactions by excluding those transfers in which the seller and the buyer have very similar names indicating that they may be the same entity and have escaped our initial disambiguation efforts.
By restricting/limiting our investigation to granted patents filed between 1990 and 2013, we create a sample of 3,515,648 patents, 19.8% of which transferred at least once. Then, we select only patents with a university as first assignee, for a total of 106,075 (3% of USPTO patents).

University patenting at USPTO: key figures

Over the past 30 years, we have assisted to a strong growth in university patent filings at the USPTO, which was due first to US institutions increasing their activity in response to the Bayh-Dole Act (up until the early 2000s) and then to the sharp rise of foreign ones (see Figure A1 in Appendix). University patents granted by the USPTO were about 2% of the total number of issued patents at the beginning of the 1990s and now they stand at around 4%.
US universities obtain the majority of university patents in our study, representing two-thirds of the sample (68,819, corresponding to 64.9% of all USPTO university patents). Other countries with more than 3,000 university patents are Japan (5645), Taiwan (4058 patents), China (3988), South Korea (3883), and Canada (3469).
Among the top universities, the University of California is the institution with more granted patents (7,537), followed by MIT (3,327), the University of Florida (3,026), Stanford University (2,374) and California Institute of Technology (2,347). Among the foreign universities, Tsinghua University is on the top (with 1287 patents), followed by the National Taiwan University (688) and the University of Hong Kong (676).
About 12.3% of these patents are transferred at least once, and in most cases they are subject to only one transfer: only 1.9% of patents are transferred more than twice. Most university patents are sold to companies (7,959 patents, representing the 7.5% of all university patents - See Table 1).
However, a few of them (326 patents, representing the 0.3% of the sample) are directly transferred to PAEs. It is noticeable that almost all the transfers to PAEs occurred in the last ten years (see Figure 1), suggesting a new rising phenomenon. About 3% of university patents are transferred to other universities. This could be due to a real transfer between two institutions, especially when university co-assignees sell their part to other university co-assignees, to reallocate shares, reflecting a false change of ownership.

Table 1. Transfers of university patents, by buyer sector

Buyer's sector	# of patents (a)	(%)	# of patents [only first transfer] (b)	(%)
PAE	383	0.4%	326	0.3%
COMPANY	8,357	7.9%	7,959	7.5%
UNIVERSITY	3,379	3.2%	2,969	2.8%
HOSPITAL	120	0.1%	114	0.1%
GOV NON-PROFIT	1,407	1.3%	1,321	1.2%
INDIVIDUAL	855	0.8%	766	0.7%
Total patents		*106,074*		
Transferred patents		*13,077 (12.3%)*		

Note. A patent is counted more than once when it has been transferred to two (or more) applicants of a different sector during its life. Column (a) shows the number of patents bought by type of buyer considering both direct (first) transfers and indirect transfers. Column (b) shows only the number of patents by type of buyer that have been *directly* transferred from universities. Years of filing: 1990-2013. Granted patents only.

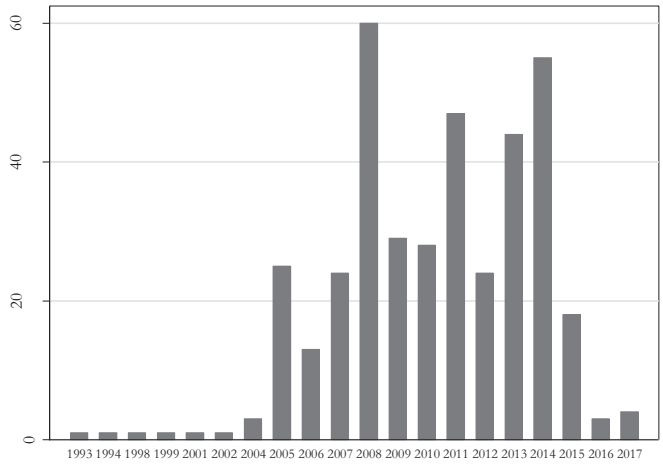


Figure 1. Transfers to PAEs, by transfer year

Note. The figure shows the number of university patents transferred to PAEs by transfer year.

Not surprisingly, most of the university patents acquired by PAEs are related to the high-tech sector. Following the international patent classification (IPC), eight macro sections, we find that more than the 80% of the transfers regard *Physics* and *Electricity*.

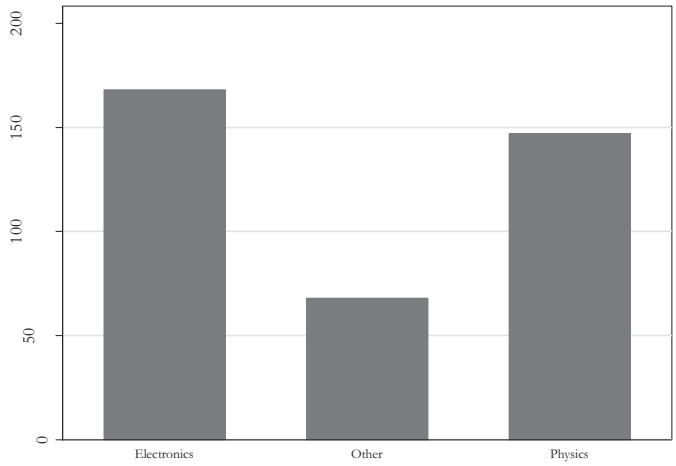


Figure 2. Transfers to PAEs, by technology

Note. IPC section ("Other" includes A = Human Necessitites, B = Performing Operations; Transporting, C = Chemistry; Metallurgy, D = Textiles; Paper, E = Fixed Constructions, F = Mechanical Engineering; Lighting; Heating; Weapons; Blasting)

Among the universities that transfer more patents to PAEs, we identify North Carolina State University (with 38 patents transferred to PAEs), Sungkyunkwan University (36), University of Texas System (24), Yonsei University (17) and Duke University (11). In total, the PAD-USPTO records indicates that 92 universities (33 of which are based in US, 13 in South Korea and 10 in UK) directly transfer patents to a PAE. Regarding PAEs as buyers of universities patents, two entities account for around 70% of transfers: Intellectual Ventures (with 188 patents, of which 161 as first transfer) and Intellectual Discovery (with 72 patents, of which 72 as first transfer) (See Table 3).

Table 2. TOP 5 Universities, by number of patents transferred to PAEs (direct transfer).

PAE	Direct Transfers
North Carolina State University (US)	38 (11.7%)
Sungkyunkwan University (KR)	36 (11.0%)
University of Texas System (US)	24 (7.4%)
Yonsei University (KR)	17 (5.2%)
Duke University (US)	11 (3.4%)

Note. Years of filing: 1990-2013. Granted patents only.

Table 3. TOP 5 PAEs, by number of university-transferred patents.

PAE	Acquired patents (a)	Acquired patents (directly from universities) (b)
INTELLECTUAL VENTURES	188 (49.3%)	161 (49.5%)
INTELLECTUAL DISCOVERY	72 (18.8%)	72 (22.2%)
TESSERA	16 (4.2%)	6 (1.9%)
RPX	16 (4.2%)	4 (1.2%)
ROCKSTAR	10 (2.6%)	0 (0.0%)
All PAEs	383	326

Note. Column (a) shows the number of patents bought by PAE as buyer considering both direct (first) transfers and indirect transfers. Column (b) shows only the number of patents by PAE that have been *directly* transferred from universities. Years of filing: 1990-2013. Granted patents only.

Characteristics of PAE-acquired university patents

In this section, we look at the characteristics of university patents acquired by PAEs to investigate whether to what extent these transfers are similar to transfers involving producing companies.

As proxy of patent quality, we consider the number of citations received in the first 3 years from the filing date (3-yr Citations), the number of claims in the patent (Claims) and the age of the patent (Age) at the time of the patent.

Figure 2 shows the average values of these variables for the different types of buyers. What does emerge from the simple descriptive statistics is that PAE-acquired are of higher quality with respect to those transferred to other types of entities. However, they are also significantly older than other university patents at the time of the transfer: on average, patents transferred to PAEs are eight years old, while those transferred to producing companies are almost 3 years younger.

Of course, to the extent that the heterogeneity in the distribution across technological fields and years is important, these statistics may be biased. To control for this possibility, we perform an econometric analysis that relates the characteristics of the patent to the probability of observing a transfer to PAEs.

In particular, we estimate two models.

In the first model (model 1), we estimate the probability that a patent is transferred to a PAE against the alternative of no-transfer: that is, we exclude from the analysis all patents transferred to other types of entities different from PAEs, so that we may investigate whether those transferred to PAEs do differ from those kept by universities.

We thus estimate the following empirical model:

$$Dummy_{PAE} = \alpha_i + \alpha_1 3 - \text{yr citations} + \alpha_2 Claims_i + X_i'\delta + \varepsilon_{i,t} \quad (1)$$

The matrix X does include dummies to control for time invariant characteristics such as technological classes, the application year and the country of the patent owner.

We estimate model 1 by using Probit models (Logit models provide similar results). We cluster standard errors at the university level to control for possible serial correlations (Bertrand et al., 2004).

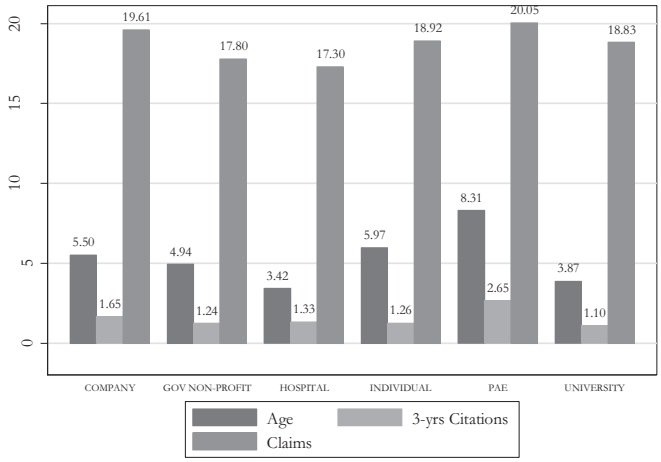


Figure 2. Characteristics of university patents, by type of buyer

Summary statistics of the variables used in the econometric exercises are reported in Appendix (Table A1). Table 4 shows the estimation results. In column (1) we do not control for baseline patent characteristics, such as technological field, application year, and country of the university. These controls are added to the specification in column (2). Controlling for observable patent characteristics, patents that receive more citations and with a larger number of claims are more likely to be transferred to PAEs rather than to remain in the university patent portfolio.

Table 4. PAEs-Acquired vs Non –Transferred University Patents

	(1) PROBIT Dummy PAE	(2) PROBIT Dummy PAE
3-yr citations	0.012***	0.015***
	(0.005)	(0.004)
Claims	0.002	0.002*
	(0.002)	(0.001)
Technological Field FEs	No	Yes
Application Year FEs	No	Yes
Country FEs	No	Yes
Observations	93320	83198
Pseudo-R2	0.0087	0.1899

Note. Cluster standard error (at the applicant/university level) in parenthesis; *** p<0.01, ** p<0.05, * p<0.1. *The dependent variable* is a dummy variable equals to one if the first buyer of the university patent is a PAE and 0 if the patent has not been transferred.

In a second model we estimate the probability that a patent is (directly) transferred to PAEs *rather than* to producing companies. We replicate the empirical analysis of model 1 but including among the regressors a new variable (age), indicating the age of the patent at the time of the transfer.

Table 5. PAEs-Acquired vs Producing Companies-Acquired University Patents

	(1) PROBIT Dummy PAE	(2) PROBIT Dummy PAE
3-yr citations	0.01**	0.01**
	(0.006)	(0.006)
Claims	-0.0006	0.002
	(0.004)	(0.002)
Age	0.065***	0.101***
	(0.025)	(0.014)
Technological Field FEs	No	Yes
Application Year FEs	No	Yes
Country FEs	No	Yes
Observations	7994	7321
Pseudo-R2	0.0509	0.2248

Note. Cluster standard error (at the applicant/university level) in parenthesis; *** p<0.01, ** p<0.05, * p<0.1. *The dependent variable* is a dummy variable equals to one if the first buyer of the university patent is a PAE and 0 if it is a producing company.

We thus estimate the following empirical model:

$$Dummy_{PAE} = \alpha_i + \alpha_1 3 - \text{yr citations} + \alpha_2 Claims_i + \alpha_3 Age_i + X_i'\delta + \varepsilon_{i,t} \qquad (2)$$

Table 5 shows the estimation results. University patents transferred to PAEs differ from those transferred to producing companies for number of citations received in the first three years from the filing date and for the age. In term of marginal effect, one year more increases the probability to observe a transfer to a PAE (instead to a producing company) by 0.009. Corroborating the results of Orsatti and Sterzi (2018) and Abrams et al. (2019), this result seems to suggest that PAEs are particularly active in the business of patent monetization, but less involved in technology transfer and intermediation activities. This is because the technology they buy is relatively old to the market and only a small share of the acquired patents (10%) is subsequently transferred to producing companies.

Conclusions

Our study provides a first extensive evidence of transfers of university patents to Patent Assertion Entities (PAEs) at the USPTO. PAEs have been accused of making new technology more expensive and of using the patent system in a way that is contrary to the purpose for its creation. For this reason, considering the role traditionally performed by universities in the production and dissemination of knowledge, transfers to PAEs alarm policy makers and academics. Not surprisingly, commentators have already highlighted a possible conflict between the stated purpose of the Bayh-Dole Act-like legislative acts and certain university

monetization strategies that make societal access to university inventions more difficult and expensive (Eisenberg & Cook-Deegan, 2019).

In our study we find that only a small share (0.3%) of university patents has been transferred to PAEs. However, most transfers occur in the last years. Two PAEs only seem to be particularly interested in university inventions (Intellectual Ventures and Intellectual Discovery), buying around the 70% of university patents. Moreover, most of the transfers occur in the high-tech sector.

PAEs target patents that are, on average, of high quality and quite old, suggesting that PAEs are particularly active in the business of patent monetization but less involved in technology transfer and intermediation activities.

Acknowledgments

All our thanks to Natalia Zinovyeva, Stuart Graham, Philippe Gorry, Gianluca Orsatti and to participants at the BRICK Workshop on The Organisation, Economics and Policy of Scientific Research (Bordeaux, 2019), BRICK - Collegio Carlo Alberto Seminar (Turin, 2019), Epip Conference (Berlin, 2018), De Paul Law Seminar (Chicago, 2019), IPSC Conference (Berkley, 2018), X IBEO Workshop (Corte, 2019).

We would like to thank Gianluca Tarasconi for his help in the data construction and for his precious suggestions, and Deivyd Alexander Velasquez Espitia for helpful research assistance. We would also like to thank the ANR (NPEIE Project, no. ANR-17-CE26-0014-01 - https://npeie.org) and the International Research Project (IRP) CNRS-IRP ALLIES "Associated Laboratory on Linkages between Innovation and Environmental Sustainability" for financial support.

Reference

Bertrand, M., Duflo, E. & Mullainathan, S. (2004). How much should we trust differences-in-differences estimates?, *The Quarterly Journal of Economics* 119(1), 249-275.

Cahoy, D. R., Kwasnica, A. M., & Lopez, L. A. (2016). The Role of Auctions in University Intellectual Property Transactions. *Duq. L. Rev*., 54, 53.

Cohen, L., Gurun, U. & Kominers, S. D. (2014). Patent trolls: Evidence from targeted firms, NBER Working Paper 20322.

Drivas, K., Lei, Z. & Wright, B.D. (2017). Academic patent licenses: Roadblocks or signposts for nonlicensee cumulative innovation?, *Journal of Economic Behavior & Organization*, 137, 282-303.

Eisenberg, R. S., & Cook-Deegan, R. (2018). Universities: The Fallen Angels of Bayh-Dole?. *Daedalus*, 147(4), 76-89.

Ewing, T. & Feldman, R. (2012). The giants among us. *Stan. Tech. L. Rev*., 1.

Firpo, T. & Mireles, M.S. (2018). Monitoring Behavior: Universities, Nonprofits, Patents, and Litigation, *SMU Law Review*, 71(2), 505.

Fischer, T. & Henkel, J. (2012). Patent trolls on markets for technology. an empirical analysis of NPEs patent acquisitions, *Research Policy*, 41(9), 1519-1533.

Fusco, S. (2013). Markets and Patent Enforcement: A Comparative Investigation of Non-Practicing Entities in the United States and Europe, *Mich. Telecomm. & Tech. L. Rev*., 20, 439.

Graham, S. J., Marco, A. C., & Myers, A. F. (2018). Patent transactions in the marketplace: Lessons from the uspto patent assignment dataset. *Journal of Economics & Management Strategy*, 27(3), 343-371.

IAM Magasin, (2019), Issue 94 March/April.

Ledford, Heidi. "Universities struggle to make patents pay." *Nature News* 501, no. 7468 (2013): 471.
Leiponen, A. & Delcamp, H. (2019). The anatomy of a troll? Patent licensing business models in the light of patent reassignment data, *Research Policy*, 48(1), 298-311.
Love, B. J. (2013). An Empirical Study of Patent Litigation Timing: Could a Patent Term Reduction Decimate Trolls Without Harming Innovators?, *University of Pennsylvania Law Review*, 1309-1359.
Merrill, S. A., & Mazza, A. M. (Eds.). (2011). Managing university intellectual property in the public interest, National Research Council, Washington, D.C.: National Academy Press.
McDonough, J. F. I. (2006). The myth of the patent troll: An alternative view of the function of patent dealers in an idea economy, Emory LJ, 56, 189.
Orsatti, G. & Sterzi, V. (2018). Do Patent Assertion Entities Harm Innovation? Evidence from Patent Transfers, Cahiers du Gretha, No. 2018-08.
Rai, A.K. & Eisenberg, R.S. (2003). Bayh-Dole reform and the progress of biomedicine. *Law & Contemp. Probs*., 66, 289.
Rooksby, J.H., (2010). University initiation of patent infringement litigation. *J. Marshall Rev. Intell. Prop*. L., 10.
Serrano, C. J. (2010). The dynamics of the transfer and renewal of patents. *The RAND Journal of Economics* 41(4), 686-708.
Thompson, N.C., Ziedonis, A.A. & Mowery, D.C. (2018). University licensing and the flow of scientific knowledge. *Research Policy*, 47(6), 1060-1069.
Kling, R. & Elliott, M. (1994). *Digital library design for usability*. Retrieved December 7, 2001 from:

Appendix

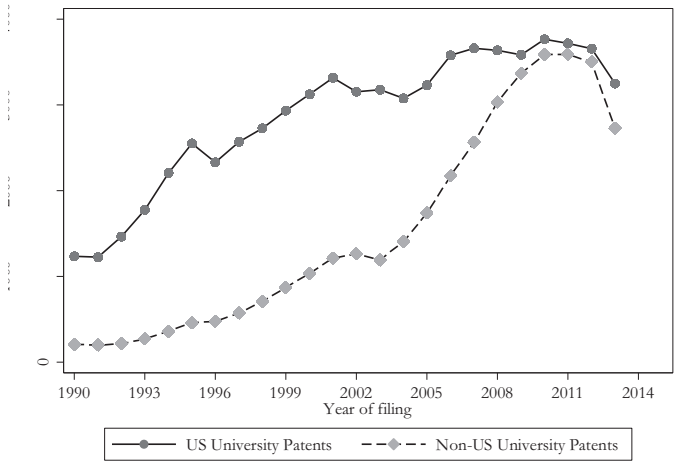


Figure A1. Number of university patents at the USPTO: US vs. Foreign Universities

Note. Number of University patents at USPTO by year of application. Co-applications with other types of sectors (COMPANY, INDIVIDUAL …) are not considered.

Table A1. Summary Statistics

	Mean	Std. Dev.	Min	Max
Model (1) - Obs. 93320				
3-yr citations	1.10	3.16	0	189
Claims	18.19	13.95	1	333
Model (2) - Obs. 7994				
3-yr citations	1.69	3.96	0	98
Claims	19.63	15.00	1	172
Age	5.62	3.96	1	20

Global country-level patterns of Mendeley readership performance compared to citation performance: does Mendeley provide a different picture on the impact of scientific publications across countries?

Rodrigo Costas[1,2] Zohreh Zahedi[2] and Juan Pablo Alperín[3]

[1] rcostas@cwts.leidenuniv.nl; z.zahedi.2@cwts.leidenuniv.nl
Centre for Science and Technology Studies (CWTS), Leiden University, Leiden (The Netherlands)
[2] DST-NRF Centre of Excellence in Scientometrics and Science, Technology and Innovation Policy, Stellenbosch University, Stellenbosch, (South Africa)

[3] juan@alperin.ca
Simon Fraser University, Vancouver (Canada)

Abstract

As bibliographic reference managers like Mendeley made their data openly available, it became possible to track where in the world research was being saved from. This data offered the opportunity to better understand how research circulates at a global scale with measures that go beyond citations. This paper explores this circulation by studying fluctuations in rankings between countries when they are based on mean normalized citation scores (MNCS) or on mean normalized Mendeley readership scores (MNRS). Results show that both indicators are moderately correlated at the country level, but that countries from the Global South (namely African and South American countries) perform better when ranked by Mendeley readership than by citations. In addition, publications from South America and Africa tend to have a lower citation impact compared to those from Europe and North America, even when compared with publications that have the same number of readers. These results suggest that the indicator chosen (i.e., citations or Mendeley readers) creates different (dis)advantages among scholarly actors (e.g. countries, research organizations, journals, etc.). It also hints at the need to establish evaluation frameworks that consider that different metrics play different roles across institutional and geographical boundaries. We conclude by proposing further ways of exploring these metrics.

Introduction

Mendeley readership have been suggested as one of the most important sources of social media metrics from several perspectives. Firstly, it is the altmetric source with the largest coverage of scientific publications (Thelwall & Sud, 2016) and with the highest correlation with citations (Costas, Zahedi, & Wouters, 2015a). Secondly, Mendeley readership and citations exhibit similar pattern in their distributions across different fields of science (Costas, Haustein, Zahedi, & Larivière, 2016). Thirdly, Mendeley readership have a stronger conceptual proximity with citation indicators, therefore arguably having a more scholarly focus than any other social media metrics (Wouters, Zahedi, & Costas, 2018). These characteristics suggests the potential of Mendeley readership being incorporated in research evaluation processes (Zahedi, 2018). As such, normalization of the number of readership by discipline (Haunschild & Bornmann, 2016) and estimating *exchange rates* for quantifying the differences between citations and Mendeley readership across fields (Costas, Perianes-Rodríguez, & Ruiz-Castillo, 2017) have been already suggested in the literature. Moreover, it is known that geographical variations in terms of coverage and density of social media metrics across different countries exists (Alperin, 2015) which can lead to biases in social media metrics at the country level (Zahedi, 2016). It has been observed that Twitter users prefer to tweet publications from other countries than their own (Zahedi & Costas, 2017) while Mendeley users seem to prefer to save (read) publications from their own country than from any other country (Thelwall & Maflahi, 2015). However, little is known about how Mendeley readership varies across countries, and how these country-level differences in readership patterns relate to citations. As such, this study aims to fill this knowledge gap by addressing the following research question: *are there differences across countries in their*

ranking positions depending on whether they are based on mean normalized citation scores (MNCS) or on mean normalized Mendeley readership scores (MNRS)?

Methodology

For this study, a total of 10,307,814 scientific publications (article and review document types) covered in the Web of Science during the period 2010 and 2017 and having a DOI were selected. Citation and Mendeley readership indicators were calculated using the most recent data available in July 2018. Country data was obtained from the author affiliation field and, as an initial exploratory approach, full counting was used. Two basic indicators based on Mendeley readership and citations (including self-citations) and field-normalized indicators have been calculated at the publication level: *ncs* (normalized citation score) and *nrs* (normalized readership score). These values have been averaged for each country, thus obtaining a *Mean Normalized Citation Score* (MNCS) and a *Mean Normalized Readership Score* (MNRS) per country. The Web of Science Subject Category classification has been used as the underlying disciplinary classification for field-normalization. Essentially, the same normalization methodology as suggested by Waltman & Van Eck. (2013) for citations and Bornmann & Haunschild (2016) for readership have been applied here. For some analyses the *Mean Citation Score* (MCS) and the *Mean Readership Score* (MRS) without field normalization have also been considered. Finally, a total of 82 countries with more than 5,000 publications in the period of analysis have been selected for further analysis. For each country, both MNCS and MNRS indicators have been first ranked by the indicator and second by number of publications in case of ties. Countries have also been classified across 6 world regions (Africa, Asia, Europe, North America, South America, and Oceania). For comparing rankings and indicators, descriptive statistical analyses have been performed using IBM SPSS version 23.

Results

The MNCS and MNRS values at the country level were compared across world regions. The box plot distributions for each country group shows that the MNRS distributions of countries from Africa and South America are sensibly higher than their MNCS (Figure 1).

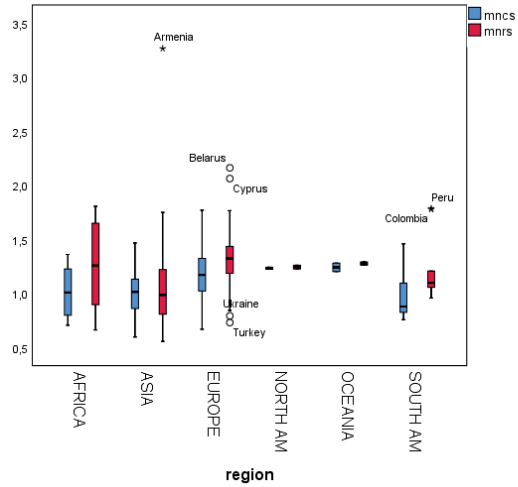


Figure 1. Distribution of countries by MNCS and MNRS indicators and by world regions

Two scatter plots showing the correlations between MNRS and MNCS and between MRS and MCS (i.e. the indicators without field normalization) reveal the relationship between readership and citations (Figure 2). Correlations of rankings are moderate in both cases (R^2 linear values of 0.395 for MNRS/MNCS and of 0.560 for MRS/MCS). This reinforces the idea that readership and citations are related, but are not totally equivalent indicators (Haustein, Bowman, & Costas, 2016).

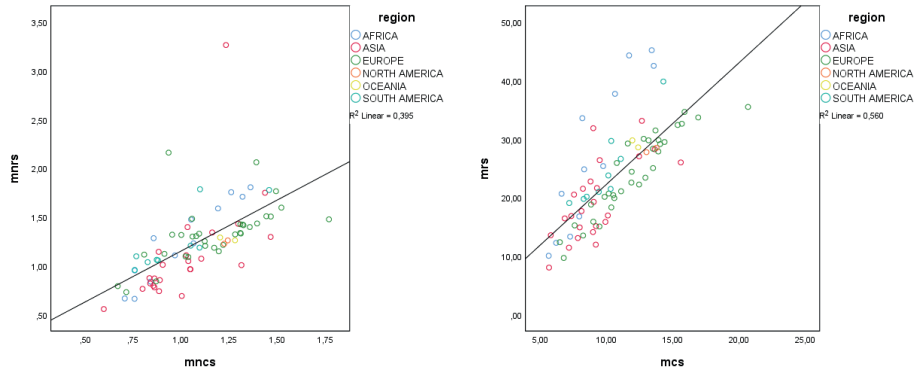


Figure 2. Scatter plot correlations between MNCS and MNRS (left graph); and MCS and MRS (right graph) for countries (colored by world regions)

As suggested by the distribution plots (Figure 1), a substantial presence of African countries lie above the regression line, suggesting that these countries have a higher position in the distribution by MNRS in relation to their position by MNCS. In order to explore this point further, the difference in rank position was calculated to measure how much countries would go up (or down) when comparing their MNCS to their MNRS. Plotting the distribution of these differences reveals very clearly how African and South American countries increase the most (more readership than citations), while the positions of European, North American (i.e. USA and Canada) and Oceania (i.e. New Zealand and Australia) countries stagnate or lose position when ranked by readership (Figure 3).

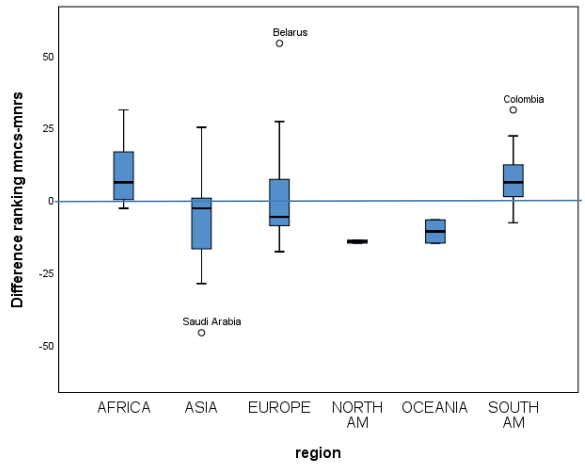


Figure 3. Ranking position difference distributions (ranking of MNCS-ranking of MNCS) for countries by world regions

Results from the above analyses highlight that countries from Africa and South America would be seen as performing better, on average, if Mendeley readership was considered in place of citations. In order to further test this idea, an *analysis of equals* is developed. For this analysis, all publications with a *nrs* value between 0.96 and 1.04 were chosen (essentially publications that are around the reference value of 1 in their fields by readership). Given this group of publications with mean readership, MNCS values were calculated for each country and averaged by world region (Figure 4).

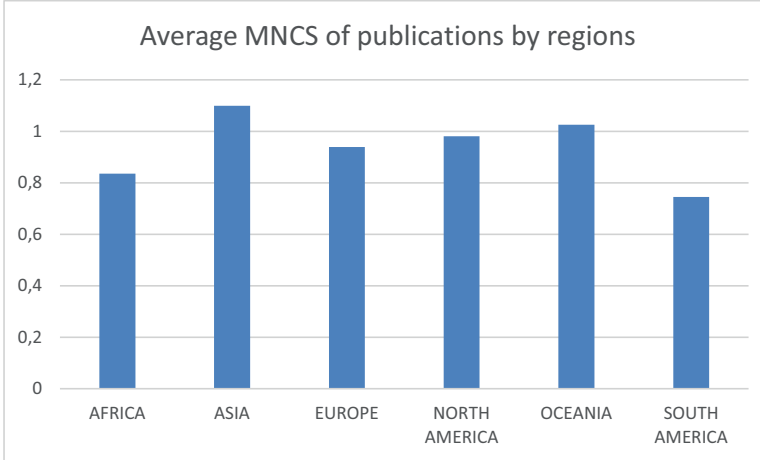


Figure 4. Analysis of publications with *nrs* values between 0.96 and 1.04 (327,355 publications) by world regions and MNCS indicator

Figure 4 shows a quite clear pattern in which publications from African and South American countries exhibit an overall lower average MNCS value. This means that for publications of equal readership impact, African and South American countries publications receive less citations.

Discussion and further research

The results presented here open new discussion of the possibilities offered by social media metrics, particularly those derived from Mendeley readership. Firstly, the metrics calculated here corroborate previous findings about the moderate correlation between citations and Mendeley readership (Sugimoto et al., 2017; Zahedi, 2018) and extend the notion that this moderate correlation is also observable at the level of countries. Secondly, the analysis presented here shows that, in spite of these moderate correlations, there are differences in how countries rank in relation to one another depending on which of the two measures are used. Interestingly, the countries from Africa and South America move from lower ranking positions when ranked by MNCS to higher ones when ranked by MNRS. Correspondingly, countries from Europe and North America move down in the rankings. This is particularly important as it sheds light on the disadvantage that countries from the Global South find themselves in when citations are given importance over readership and the circulation of the research.

The results open the relevant question of *what are the possible explanatory factors of the Mendeley readership advantage for publications from the Global South?* More specifically, it

begs the question: *why is there variation in the citation performance across countries for publications with similar levels of readership?* These effects may be the result of the Matthew effect (Merton, 1968), in which authors, read publications indiscriminately, but at the time of choosing what to cite they choose those coming from the most reputed institutions and countries, thus reinforcing the citation cumulative advantage of already scientifically advanced countries. Alternatively, it may be that this is the combined effect of the national readership affinity of Mendeley users (Thelwall & Maflahi, 2015) as well as the national affinity of citations (Sugimoto, Gong, & Larivière, 2018). That is, authors from the Global South read (and potentially would cite) more the publications of their own countries, but may be facing stronger difficulties in publishing their work in Web of Science-index journals (or just international journals). As a result, the publications they are reading from their own countries would end up having fewer citations, thereby creating a disadvantage for the publications of their own countries. Additional potential explanations may be related to the different uptake of Mendeley across countries. For example, although in this paper we have accounted for field-differences, different thematic profiles may be at play (e.g. countries with greater uptake among fields with stronger focus on social sciences, which are the fields with traditionally higher readership scores; Costas, Zahedi, & Wouters, 2015b). Or, different national preferences on how Mendeley is used (e.g. users in some countries may use Mendeley more for reading or self-education in contrast to others where it is used to curate a bibliography for citing). These are all aspects that should be explored in future research, together with the incorporation of other more technical aspects, such as the exclusion of self-citations and self-readership and the fractional counting of country authorship. Our future research will focus on corroborating the results presented in this study by tackling these issues as well as using larger datasets and expanding the set of indicators.

Even in this preliminary form, the results presented suggest that more exploration is needed about to understand *what is the potential role of Mendeley readership for research evaluation?* As suggested in this study, there may be different (dis)advantages among scholarly actors (e.g. countries, research organizations, journals, etc.) depending on the indicator chosen. It is therefore important to open a debate about how to combine these different metrics (and the perspectives they bring) into the research evaluation. This may once again raise the idea of the potential need of sort of "exchange rates" (Costas, Perianes-Rodríguez, & Ruiz-Castillo, 2017). More pragmatically, it raises the need of establishing evaluation frameworks in which both citation and readership metrics can inform different forms of impact and reception among different communities and geographical boundaries.

Acknowledgements

Rodrigo Costas is partially funded by the South African DST-NRF Centre of Excellence in Scientometrics and Science, Technology and Innovation Policy (SciSTIP).

References

Alperin, J. P. (2015). Geographic variation in social media metrics: an analysis of Latin American journal articles. Aslib Journal of Information Management, 67(3), 289–304.

Bornmann, L., & Haunschild, R. (2016). Normalization of Mendeley reader impact on the reader- and paper-side : A comparison of the mean discipline normalized reader score (MDNRS) with the mean normalized reader score (MNRS) and bare reader counts. Journal of Informetrics, 10(3), 776–788. http://doi.org/10.1016/j.joi.2016.04.015

Costas, R., Zahedi, Z., & Wouters, P. (2015a). Do "altmetrics" correlate with citations? Extensive comparison of altmetric indicators with citations from a multidisciplinary perspective. Journal of the Association for Information Science and Technology, 66(10), 2003–2019. http://doi.org/10.1002/asi.23309

Costas, R., Zahedi, Z., & Wouters, P. (2015b). The thematic orientation of publications mentioned on social media Large-scale disciplinary comparison of social media metrics with citations. Aslib Journal of Information Management, 67(3), 260–288. http://doi.org/10.1108/AJIM-12-2014-0173

Costas, R., Haustein, S., Zahedi, Z., & Larivière. (2016). Exploring paths for the normalization of altmetrics: applying the Characteristic Scores and Scales. In altmetrics16. Moving beyond counts: integrating context (pp. 1–8). Bucharest.

Costas, R., Perianes-Rodríguez, A., & Ruiz-Castillo, J. (2017). On the quest for currencies of science: field " exchange rates " for citations and Mendeley readership. Aslib Journal of Information Management. http://doi.org/10.1108/ AJIM-01-2017-0023

Haunschild, R., & Bornmann, L. (2016). Normalization of Mendeley reader counts for impact assessment. Journal of Informetrics, 10(1), 62–73. http://doi.org/10.1016/j.joi.2015.11.003

Haustein, S., Bowman, T. D., & Costas, R. (2016). Interpreting Altmetrics: Viewing Acts on Social Media through the Lens of Citation and Social Theories. In C. R. Sugimoto (Ed.), Theories of Informetrics and Scholarly Communication (pp. 372–406). Berlin, Boston: De Gruyter. http://doi.org/10.1515/9783110308464-022

Larivière, V., Gong, K., Sugimoto, C.R. (2018). Citations strength begins at home. Nature, 564, S70-S71. https://doi.org/10.1038/d41586-018-07695-1

Merton, R. K. (1968). The Matthew Effect in Science. Science, 159(3810), 56–63

Sugimoto, C. R., Work, S., Larivière, V., & Haustein, S. (2017). Scholarly use of social media and altmetrics: A review of the literature. Journal of the Association for Information Science and Technology, 68(9), 2037–2062. http://doi.org/10.1002/asi.23833

Thelwall, M., & Maflahi, N. (2015). Are scholarly articles disproportionately read in their own country? An analysis of Mendeley readers. Journal of the Association for Information Science and Technology, 66(6), 1124-1135.

Thelwall, M., & Sud, P. (2016). Mendeley readership counts: An investigation of temporal and disciplinary differences. Journal of the Association for Information Science and Technology, 67(12), 3036–3050. http://doi.org/10.1002/asi.23559

Waltman, L., & Van Eck, N.J. (2013). Source normalized indicators of citation impact: an overview of different approaches and an empirical comparison. Scientometrics, 96(3), 699-716. doi:10.1007/s11192-012-0913-4

Wouters, P., Zahedi, Z., & Costas, R. (2018). Social media metrics for new research evaluation. In W. Glänzel, H. F. Moed, U. Schmoch, & M. Thelwall (Eds.), Handbook of Quantitative Science and Technology Research. Springer.

Zahedi, Z. (2016). Could we start to talk about an 'altmetric divide'? On the imbalance in the access and use of social media platforms across countries [Blog Post]. 3:AM conference, 27-29 September 2016, Bucharest, Romania. Retrieved from: https://goo.gl/QTGmrD

Zahedi, Z., & Costas, R. (2017). How visible are the research of different countries on WoS and Twitter? an analysis of global vs. local reach of WoS publications on Twitter. In 16th International Conference on Scientometrics & Informetrics (ISSI), Wuhan, China. doi (Vol. 10, p. m9).

Zahedi, Z. (2018). Understanding the value of social media metrics for research evaluation (Doctoral dissertation). Supervisors: Wouters P. & Costas R. Leiden University, The Netherlands.

Identifying communities of interest in social media: Microbiology as a case study

Wenceslao Arroyo-Machado[1], Daniel Torres-Salinas[2] and Nicolas Robinson-Garcia[3]

[1] *wences91@gmail.com*
International School for Postgraduate Studies, University of Granada, Granada (Spain)

[2] *torressalinas@gmail.com*
Vice-Rector's Office for Scientific Policy and Research, EC3metrics SL, University of Granada, Granada (Spain)

[3] *elrobinster@gmail.com*
Delft Institute of Applied Mathematics, TU Delft (Netherlands)

Abstract

As part of altmetrics, social media metrics focuses his attention in the quantitative analysis of mentions from social media to scientific literature. One of the approaches to analysing the interactions produced in social media is to identify discussion topics and the actors involved in it, missing more research focusing in the link between them. The main goal of this paper is to propose a methodology that make possible the detection of research topics and to whom can be of interest either within or outside the scientific realm. To do it we combine users' networks and overlay mapping of topics to visualize and identify communities of attention, allowing to contextualize the mentions to scientific publications. The dataset used is formed by the union of publications in the field of Microbiology and the mentions received from news media, policy briefs and, mostly, twitter.

Introduction

Altmetrics emerged as a direct response to traditional citation-based indicators. It presented itself as an alternative way to analyse the impact of scientific work (Priem et al., 2010). Since then, there is still much debate as to its capabilities and even as to what it actually encompasses and how it should be named (Ronald & Fred, 2013). As a means of clarification, some authors refer to social media metrics as an area which strictly focuses on the quantitative analysis of mentions to scientific literature from social media (Haustein et al., 2014; Wouters, Zahedi & Costas, 2019). Within the different social media platforms, Twitter, along with Mendeley, offers the most extensive and intense coverage of scientific publications (Robinson-Garcia et al., 2014; Costas, Zahedi & Wouters, 2015). This platform allows user to interact and disseminate scientific work in different ways; either by quoting a paper through retweets, responding to others or supporting (likes) the tweets of others. By doing so, it is expected that by tracking mentions to publications in this platform, we will be able to identify discussions around scientific topics. However, this dialog is in many cases blurred by bots (Haustein et al., 2016) or, -partly due to limitation on the number of characters messages can have-, bot-like behaviour from individuals (Robinson-Garcia et al., 2017). These limitations affect the capacity to interpret any type of quantitative indicators based on Twitter activity, discarding the effectiveness of these approaches towards the study of scholarly dissemination in Twitter.

As a result, there has been a change of perspective and shift towards the use of network analysis and mapping techniques, changing also the goal for which Twitter is used (Robinson-Garcia, van Leeuwen & Rafols, 2018). Network analyses and visualization techniques serve as descriptive representations on interactions between actors consuming scientific literature or interacting with scholars. Several methodologies have been proposed to visualize the interactions between actors and scientific publications, highlighting their potential to identify communities based on social media coupling and co-social mediation (Costas, de Rijcke &

Marres, 2017), as well as engagement and exposure of users (Haustein, Bowman & Costas, 2015). Visualization techniques have also been applied to identify topics of discussion in social media by producing thematic landscapes (Wouters, Zahedi & Costas, 2019). Moreover, it has been suggested that one should be able to highlight patterns of interest or engagement of societal actors in different research topics by overlaying scores from altmetric sources on science maps (Noyons & Ràfols, 2018).

The approaches described above analyse either the subject of discussion (scientific literature) or the actors involved in such 'discussion'. However, few studies have attempted to link between these two groups (although this is suggested in Costas, 2017; Wouters, Zahedi & Costas, 2018). Costas, de Rijcke & Marres (2017) is to our knowledge, the only other study in which they do so by applying what they referred to as 'heterogeneous couplings' as a means by which actors and objects can be linked through 2-mode networks. In this paper, we propose the combination of users' networks and overlay mapping of topics as a means to visualize and identify communities of attention in Twitter, news media and policy briefs. The goal is to bridge between research areas and actors. By doing so, we can contextualize Twitter mentions to publications and areas which are prone to be of interest to specific stakeholders.

Here we present our first steps on building a methodology by which we can better understand which research topic and to whom can be of interest either within or outside the scientific realm. Following up from a recent study (Robinson-Garcia, Arroyo-Machado & Torres-Salinas, 2019) in which we mapped discussions in Twitter, news media and policy briefs surrounding research in the field of Microbiology, we expand our analysis by identifying and characterising users with discussions.

Materials and Methods

In October 2018, we retrieved a total of 382,998 records of publications indexed in the subject categories of Microbiology and Biotechnology & Applied Microbiology from Web of Science in the 2012-2018 period. We then identified which of these were covered by Altmetric.com. To query Altmetric.com we need to use to use DOIs, which means that only 88.2% of our dataset could be analysed. Out this dataset a total of 174,799 where mentioned at least mentioned once by any of the sources covered by Altmetric.com.

We created a second dataset of 1,594,856 records which included mentions to the first dataset from Twitter, news media and policy briefs, among other sources. This second dataset was filtered to these three sources and linked to the first one through articles' DOI. This allowed us to build a two-mode network of mentions and publications composed of 378,898 nodes and 1,252,822 edges. For visualization purposes, we reduced the network to the main component and using a threshold of at least 10 mentions received by publication.

This two-mode network identifies all subject in the network, that is publications, Twitter accounts, news media and organizations producing policy briefs. To identify communities within this network we use the Louvain method (Blondel et al., 2008), taking into account the edge weights, and we obtain a modularity value of 0.5. In this way, each cluster is composed by social media actors and scientific publications among which there are strong and common mentions. We identify two types of relations of similarity between them: social media coupling and co-social media-tion (Costas, de Rijcke & Marres, 2017). Table 1 includes a description of this network and its clusters detected. Secondly, we aim at identifying the most discussed topics in some of the identified communities. For this, we create a thematic landscape based on terms contained in the titles of all articles with mentions in Altmetric.com,

offering a basemap of what is discussed in social media overall. This map is created using VOSViewer, which applies a modularity-based community detection algorithm (Waltman and Eck 2013). The clusters generated by VOSViewer were later corroborated with experts in microbiology (for more details on this we refer to Robinson-Garcia, Arroyo-Machado & Torres-Salinas, 2019). We finally overlay for each of the communities detected in the two-mode network, the terms of the publications mentioned by each community.

Table 1. Descriptive of the 15 communities detected and the documents mentioned within them in Altmetric.com by Twitter, news media and policy documents from all publications indexed in Web of Science subject categories Microbiology and Biotechnology & Applied Microbiology for the 2012-2018 period. In bold clusters used in this paper.

Cluster	Twitter actors (tweets)	News story actors (mentions)	Policy document actors (mentions)	Papers (mentions received)	Total internal mentions	Total actors (mentions)
0	**2836 (131707)**	**5 (139)**	**0 (0)**	**5205 (138728)**	**89768**	**8046 (131846)**
1	**2794 (113782)**	**1 (87)**	**0 (0)**	**4934 (109117)**	**81277**	**7729 (113869)**
9	1918 (93533)	1 (16)	0 (0)	3802 (94906)	53617	5721 (93549)
8	1953 (69608)	5 (57)	1 (22)	3439 (64152)	36904	5398 (69687)
11	759 (13113)	675 (45292)	6 (405)	2291 (61919)	32870	3731 (58810)
2	2366 (58248)	19 (490)	4 (38)	2276 (58235)	34821	4665 (58776)
5	**1367 (45124)**	**11 (827)**	**8 (546)**	**2343 (45516)**	**33143**	**3729 (46497)**
7	548 (13964)	4 (225)	0 (0)	632 (13696)	6786	1184 (14189)
14	14 (120)	102 (8202)	0 (0)	179 (9567)	6457	295 (8322)
6	115 (1984)	2 (33)	0 (0)	112 (1522)	843	229 (2017)
13	38 (1016)	0 (0)	0 (0)	72 (813)	531	110 (1016)
12	4 (138)	0 (0)	0 (0)	10 (240)	13	14 (138)
3	5 (118)	0 (0)	0 (0)	25 (424)	30	30 (118)
4	2 (6)	0 (0)	0 (0)	1 (4)	2	3 (6)
10	1 (1)	0 (0)	0 (0)	1 (2)	1	2 (1)
Total	*14720 (542462)*	*825 (55368)*	*19 (1011)*	*25322 (598841)*	*377063*	*40886 (598841)*

Results

We identified a total of 15 clusters in the main component of a two-mode network of publications, Twitter accounts, news media organizations and policy organizations. Table 1 offers basic statistics on each of these clusters. The main component includes 10% of the nodes. We do this by selecting only those nodes with a minimum indegree 10 or minimum outdegree 10, which is equivalent to receiving or mentioning at least 10 publications. Still, the network includes 48% of the mentions. On average each actor performs 38 different mentions (±97.02), while publications receive an average of 24 different mentions (±31.64). Twitter emits most of the mentions, representing 36% of the nodes of the network (around 50% in the full network) and 91% of the mentions observed. Publications represent 62% of the nodes. Policy briefs' mentions are the least appreciable in the network, representing less than 1%, while mentions from news media represent approximately 9%.

Figure 1 shows this network, first distinguishing by type of node (1A) and then by cluster (1B). We observe that news media represent two different clusters. Cluster 11 is the only one composed of a mixture of twitter and news media, although the latter dominate in number of mentions, while cluster 14, with a majority of news media, is clearly separated from the main

network. Except for those two clusters, all have a disproportionate proportion of Twitter users and mentions compared to the other two types of nodes (news media and policy organizations).

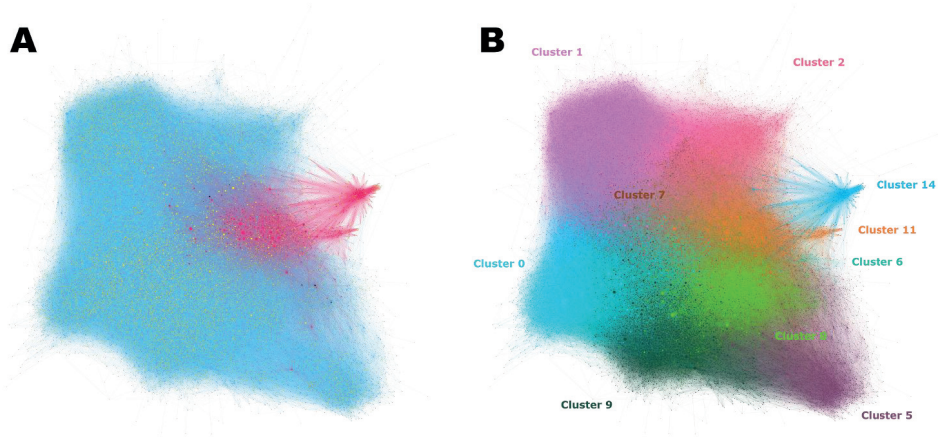


Figure 1. Giant component of two-mode top 10% network of Microbiology publications and altmetric actors mentioning them. A nodes in yellow are publications, in blue are Twitter accounts, in red are news media and in black, organizations producing policy briefs; B nodes colored by cluster.

Following, we use a term map constructed by Robinson-Garcia, Arroyo-Machado & Torres-Salinas (2019). This map is created using term from titles of all microbial publications indexed in Altmetric.com. Figure 2 shows the basemap/with the seven large topics where nodes represent words and noun phrases from titles and colors represent clusters. This map is then used to overlay and identify in which topics are discussions focused by each cluster. Here we will focus on three clusters or communities: cluster 0, cluster 1 and cluster 5. The first two are part of the main part component of the network while the last one combines Twitter accounts, news story and policy documents.

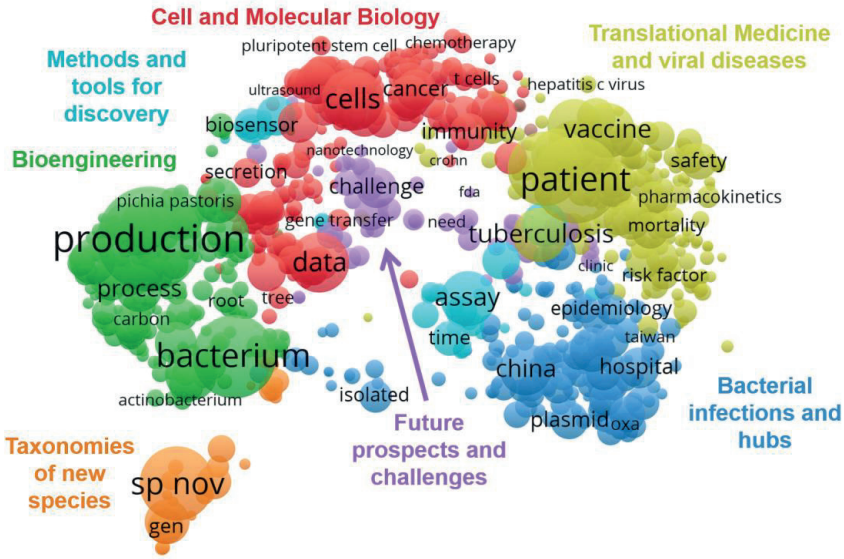


Figure 2. Top 60% most relevant terms occurring in titles of publications included in Altmetric.com from the Web of Science subject categories of Microbiology and Biotechnology & Applied Microbiology. Minimum threshold: 50 co-occurrences. Source: Robinson-Garcia, Arroyo-Machado, Torres-Salinas (2019)

Figure 3, 4 and 5 overlays these three clusters on the term map. Cluster 0 focuses mostly on bioengineering and also, but to a lesser extent, some areas immediately close to it, especially taxonomies of new species, future prospects and challenges and cell and molecular biology (Figure 3).

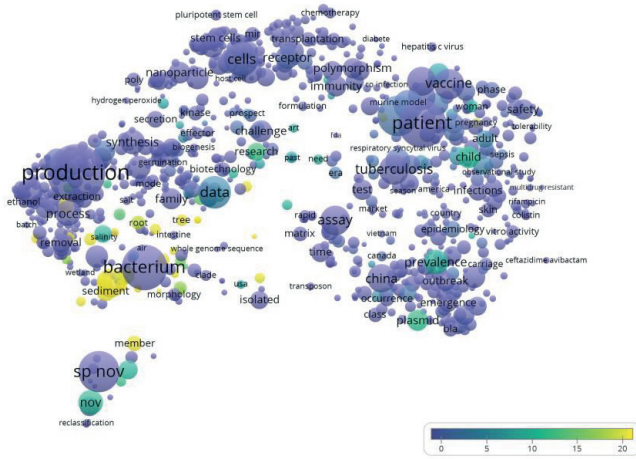


Figure 3. Overlay terms map for cluster 0 mentions to microbial literature based on figure 2. Color: normalized mentions to terms made by the cluster.

Cluster 1 discusses mainly topics related to cell and molecular biology. Out of this cluster there are two outstanding terms: literature inside the topics of translational medicine and viral diseases, and science inside future prospects and challenges (Figure 4).

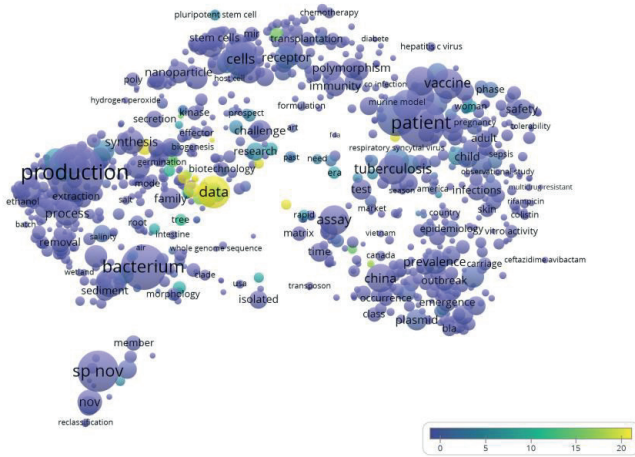


Figure 4. Overlay terms map for cluster 1 mentions to microbial literature based on figure 2. Color: normalized mentions to terms made by the cluster.

Finally, cluster 5 is focused mainly on translational medicine and viral diseases and bacterial outbreaks. It covers a greater number of terms, reaching some bordering terms on future prospects and challenges (Figure 5).

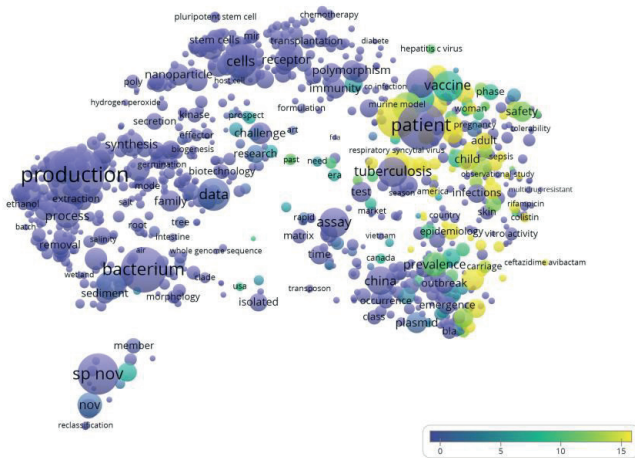


Figure 5. Overlay terms map for cluster 5 mentions to microbial literature based on figure 2. Color: normalized mentions to terms made by the cluster.

Table 2. Highlights of the three cluster studied with the main factors which contextualize these different communities

	Cluster 0	Cluster 1	Cluster 5
CHARACTERISTICS AT THE CLUSTER LEVEL			
Cluster structure	35.3% - Twitter accounts 0.1% - News media 64.7% - Publications	36.2% - Twitter accounts 0.0% - News media and 63.8% - Publications	36.7% - Twitter accounts 0.3% - News media 0.2% - Policy briefs 62.8% - Publications
Coverage by actor for the complete set of pubs	Twitter - 30.6% pubs News media - 0.11% pubs	Twitter - 29.3% pubs News media - 0.1% pubs	Twitter - 16.7% pubs News media - 1.8% pubs Policy briefs - 1.4% pubs
Most discussed topic from term map	Bioengineering	Cell and molecular biology	Translational medicine and viral diseases and bacterial infections and hubs
Most discussed terms from term map	Tree, taxonomy and human gut	Human genome, literature and genome assembly	Infectious diseases society, america and update
CHARACTERISTICS BASED ON THE TOP 25 SOCIAL MEDIA ACTORS			
Publication coverage of social media actors	19.56% of the publications of the entire network are mentioned	17.59% of the publications of the entire network are mentioned	12.46% of the publications of the entire network are mentioned
Types of Twitter accounts	13 bots, 9 academics, 2 scientific communities and 1 journal	14 bots, 7 academics, 2 professionals, 1 company and 1 journal	8 bots, 10 professionals, 2 academics, 2 journals, 1 hospital, 1 society and 1 specialized website
Types of news media	3 specialized media, 1 university press and 1 scientific association	1 journal press	7 specialized media, 2 journals and 1 research centre
Types of policy organizations	-	-	3 government organizations (national), 5 health organizations (4 national and 1 global)
Most discussed topic from term map	General interest in the field of microbiology	Mostly mentioning publications in the field of bioinformatics	Actors are related to hospitals and clinical medicine. Focused on viral diseases and bacterial infections
Most influential actors	@jcamthrash with 5441 mentions	@yeast_papers with 6147 mentions	@AntibioticResis with 6496 mentions

Next, we characterize these three clusters (Table 2). We do so by looking at global metrics and analysing the top 25 most active actors for each cluster, differentiating by type. We first a significant presence of Twitter bots among the top 25 Twitter users: cluster 0 has 13, cluster 1 has 14 and cluster 5 has 8. Referred to this top, all are composed of Twitter accounts with the exception of the last, in which there is a news media and a policy brief. These accounts have been identified as hubs since they cover a large percentage of publications from the entire network: top 25 most active actors of cluster 0 mentions 19.56%; cluster 1, 17.59%; and cluster 5, 12.46%. Also, the news media of the clusters are mostly specialized media, while the policy briefs, only present in cluster 5, are almost all national institutions, highlighting

those focused on the health field. The analysis of these top accounts shows a general approach to microbiology in cluster 0, while cluster 1 has a clear orientation to the bioinformatics profile, while the latter is even clearer its focus on the hospital environment and bacterial infections, with numerous professionals, specialized media on pharmacology and disease control organizations.

Concluding remarks

This paper attempts to expand our understanding on how and who discusses literature in social media. We present a methodological framework by which actors and topics of discussion are linked to each other by combining two-mode networks of actors with overlay mapping of topics. Building on a previous study (Robinson-Garcia, Arroyo-Machado & Torres-Salinas, 2019), we use a set of microbial publications as case study. We then identify different clusters or communities of actors discussing microbial literature and overlay the publications discussed by each cluster on a global term map of topics discussed on Microbiology. In this paper we have just focused on the analysis of the 3 of the 15 identified communities. We expect to further extent our analysis to all 15 communities as well as add other altmetric data sources (e.g., Mendeley, Wikipedia) to further understand the relation between the source and the communities of discussion. Also we plan to analyse different set of microbial literature by considering citation-based indicators. The aim is to answer questions such as: Are there specific communities discussing high impact research? How do disparities between citation impact and altmetrics distributed among communities?

Acknowledgements

Nicolas Robinson-Garcia is a Marie Sklodowska Curie Experienced Researcher in the LEaDing Fellows COFUND programme.

References

Blondel, V. D., Guillaume, J.-L., Lambiotte, R. & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, *2008*(10), P10008. Retrieved from http://stacks.iop.org/1742-5468/2008/i=10/a=P10008?key=crossref.46968f6ec61eb8f907a760be1c5ace52

Costas, R. (2017). Towards the social media studies of science: social media metrics, present and future. *Bibliotecas. Anales de Investigación*, *13*(1), 1–5. Retrieved January 9, 2019, from http://arxiv.org/abs/1801.04437

Costas, R., de Rijcke, S. & Marres, N. (2017). Beyond the dependencies of altmetrics: Conceptualizing 'heterogeneous couplings' between social media and science. *The 2017 Altmetrics Workshop*. Retrieved from http://altmetrics.org/wp-content/uploads/2017/09/altmetrics17_paper_4.pdf

Costas, R., Zahedi, Z. & Wouters, P. (2015). Do "altmetrics" correlate with citations? Extensive comparison of altmetric indicators with citations from a multidisciplinary perspective. *Journal of the Association for Information Science and Technology*, *66*(10), 2003–2019. Retrieved from https://arxiv.org/abs/1401.4321

Haustein, S., Bowman, T. D. & Costas, R. (2015). "Communities of attention" around scientific publications: who is tweeting about scientific papers? *Social Media and Society 2015 International Conference.* Toronto. Retrieved from https://www.slideshare.net/StefanieHaustein/communities-of-attention-around-journal-papers-who-is-tweeting-about-scientific-publications

Haustein, S., Bowman, T. D., Holmberg, K., Tsou, A., Sugimoto, C. R. & Larivière, V. (2016). Tweets as impact indicators: Examining the implications of automated "bot" accounts on Twitter. *Journal of the Association for Information Science and Technology*, *67*(1), 232–238. John Wiley & Sons, Ltd. Retrieved from https://doi.org/10.1002/asi.23456

Haustein, S., Larivière, V., Thelwall, M., Amyot, D. & Peters, I. (2014). Tweets vs. Mendeley readers: How do these two social media metrics differ? *it - Information Technology*. Retrieved from https://www.degruyter.com/view/j/itit.2014.56.issue-5/itit-2014-1048/itit-2014-1048.xml

Noyons, E. & Ràfols, I. (2018). Can bibliometrics help in assessing societal contributions of agricultural research? Exploring societal interactions across research areas. *23rd International Conference on Science and Technology Indicators (STI 2018), September 12-14, 2018, Leiden, The Netherlands*. Centre for Science and Technology Studies (CWTS).

Priem, J., Taraborelli, D., Groth, P. & Neylon, C. (2010). Altmetrics: A manifesto. Retrieved from http://altmetrics.org/manifesto/

Robinson-Garcia, N., Arroyo-Machado, W. & Torres-Salinas, D. (2019). Mapping social media attention in Microbiology: Identifying main topics and actors. *FEMS Microbiology Letters*, submitted.

Robinson-Garcia, N., Costas, R., Isett, K., Melkers, J. & Hicks, D. (2017). The unbearable emptiness of tweeting—About journal articles. (W. Glanzel, Ed.)*PLOS ONE*, *12*(8), e0183551. Public Library of Science. Retrieved January 10, 2019, from https://dx.plos.org/10.1371/journal.pone.0183551

Robinson-Garcia, N., Torres-Salinas, D., Zahedi, Z. & Costas, R. (2014). New data, new possibilities: Exploring the insides of Altmetric.com. *El professional de la información*, *101*(2), 1309-1324. Retreived from https://doi.org/10.1007/s11192-014-1263-1.

Robinson-Garcia, N., van Leeuwen, T. N. & Ràfols, I. (2018). Using altmetrics for contextualised mapping of societal impact: From hits to networks. *Science and Public Policy*, *45*(6), 815–826. Retrieved from http://dx.doi.org/10.1093/scipol/scy024

Ronald, R. & Fred, Y. Y. (2013). A multi-metric approach for research evaluation. *Chinese Science Bulletin*, *58*(26), 3288–3290. Retrieved from https://doi.org/10.1007/s11434-013-5939-3

Waltman, L. & Van Eck, N. J. (2013). A smart local moving algorithm for large-scale modularity-based community detection. *The European Physical Journal B*, *86*(11), 471. Springer.

Wouters, P., Zahedi, Z. & Costas, R. (2019). Social media metrics for new research evaluation. In W. Glanzel, H. F. Moed, U. Schmoch, & M. Thelwall (Eds.), *Springer Handbook of Science and Technology Indicators*. Springer International Publishing. Retrieved January 10, 2019, from https://arxiv.org/abs/1806.10541

Novelty as Recombination of Knowledge

Martina Iori [1] and Magda Fontana [2]

[1] *martina.iori@unito.it*
Department of Economics and Statistics University of Turin

[2] *magda.fontana@unito.it*
Despina Big Data Lab, University of Turin
Department of Economics and Statistics University of Turin

Abstract

Recombination of knowledge is commonly identified as the source of novelty in patents and in scientific articles. We introduce a new indicator of novelty as recombination of concepts in documents based on Latent Dirichlet Allocation and Hellinger distance and discuss its properties. In comparison with the measures adopted in literature, our indicator has several desirable properties: it is a continuous measure – i.e. it allows to identify breakthrough and to rank articles and patents –, it is robust to the number of topics and to the introduction of neologisms.

Introduction

The literature of novelty as recombination of concepts in documents is developed mainly in the analysis of patents. Kaplan and Vakily (2015) investigate breakthrough innovations through the analysis of patent texts. They classify patents according to their content, which is summarized by a set of topics occurring in their text. Topics are defined through Latent Dirichlet Allocation (LDA). The emergence of new topics signals innovation and topic-originating patents are labeled as breakthroughs. A limitation of this approach lies in the specific purpose of the article: the detection of breakthroughs. Since the authors aim at identifying topic-originating patents, their outcome is sensitive to the chosen number of topics. In practice, a higher number of topics will necessarily result in a larger set of novel patents. This issue is crucial because the literature on LDA is not yet able to provide indications on the optimal number of topics in a corpus thereby making the choice heuristic. Moreover, their measure is binary (breakthrough or non-breakthrough) and thus cannot generate a ranking of articles and patents.

Similar issues are present in the article by Kelly et al. (2018). They have proposed a continuous measure of breakthrough innovation based on the definition of a similarity between patents, where texts are analyzed starting from co-occurrences of words. Patents are close if the co-occurrence of words in their texts is similar and are defined as breakthrough when they are simultaneously distant (on average) from the previous set of knowledge and close to the following patented knowledge. Computing the distance between the content of a patent and the entire previous knowledge, this measure of novelty captures the average conventionality of the knowledge embedded in the patent more than its novelty. It underestimates the novelty of patents in large (conventional) sectors and underestimates the novelty of patents in small (unconventional) sectors. Moreover, Kelly et al. (2018)' novelty includes in the definition the patents impact and, consequently, captures only successful novel patents.

We introduce a measure that overcomes the limitations of both indicators in that it is not sensitive to the number of topics and allows the complete ranking of documents according to

their novelty. Novelty is computed as the distance between the distribution of topics within a document and the distributions of topics within its neighboring documents.

Measuring novelty

We define a measure of novelty in recombination of concepts based on the distance between the concepts enclosed in documents, assigning higher novelty scores to documents that are more distant from previous neighboring combinations.
More precisely, we firstly create a measure of (dis)similarity between documents and secondly we use such indicator to assign a value of novelty to each document.

Through the joint analysis of the occurrence of terms in a single document and in an entire collection of texts, text-mining techniques are designed to infer the subjects discussed in the corpus of documents. The detection of similar patterns in the co-occurrence of words or topics in different texts is used to define a measure of similarity among documents. The easiest text-mining technique is Bag of Words (BoW), a methodology that associates to each article the list of the words used in the text, paired with their frequency in the document. BoW assumes that the most frequent words are the most important ones in the definition of the article content. BoW allows obtaining a simple classification of documents, but the subsequent comparison between text contents can be very noisy. A first issue derives from the biases due to the different total occurrence (size) of words in the corpus and the different lengths of documents. This problem can be solved by using the TF-IDF (Term Frequency - Inverse Document Frequency) statistics to normalize the term frequencies. However, several issues are still unsolved. A comparison among word frequencies, normalized or not, does not account for polysemy, i.e. the different meanings of the same word in different contexts, and for the unobserved connections between terms and documents.

Since we are interested in capturing the concepts embedded in each document, a more precise approach can be obtained replacing word frequency with topics, using a topic modeling technique. Among the large number of topic modeling techniques, we focus on the Latent Dirichlet Allocation (LDA) (Blei 2003). LDA is a Bayesian probabilistic topic modeling that exploits the co-occurrences of words across texts to derive the thematic structure of the database. It infers the latent (unobserved) topics across the corpus, the distribution of words in each topic and the importance of every topic in each document .
A significant advantage of the LDA is that it assigns every topic to every document, paired with a weight proportional to the estimated relevance of the topic in the text. Those weights can be organized in vectors and then collected in a matrix of documents-topics. This many-to-many relationship is analogous to the one obtained with BoW or TF-IDF between documents and words.

We can then identify content of each document by a probability distribution over the topics and we can compare these distributions to define a similarity among the subjects treated in the documents. Since the objects of comparison are probabilities and not frequencies (as it happened in BoW), the Hellinger distance is preferable to cosine similarity. Hellinger distance measures the distance between two probability distributions, and returns 0 when two distributions over topics are disjointed and 1 if they completely overlap.

The definition of a similarity between documents is the first step in the detection of novelty. Every article is compared to the articles published in the previous n years and, since science and language evolve over time, a specific LDA is run for every n-years sample of articles in order to obtain precise and meaningful topics (see Figure 1, step 1).

The number of topics chosen to define a sample of documents, similarly to the level of other classification systems (e.g. the digit of classification codes), partially influences the computation of their similarity. Generally, the number of topics is reduced to the minimum to increase their ease of interpretation. However, in this case, we are interested in detecting the more subtle differences between the subjects of articles, more than in interpreting the meaning of topics, and we prefer to define a high number of topics for each sample. This choice is useful also to stabilize the measure and assure its robustness.

The topic modeling and the subsequent computation of the Hellinger distance between the topic probability distributions of documents pairs for every article in the previous n-years allow locating the content of each document with respect to the previous contributions (see Figure 1, step 2).
A common approach is to define novelty as the average distance from the entire previous corpus of knowledge, however if this is the case, the measure may return a misleading values of document novelty: high distance from the average content of the corpus of knowledge may be influenced by the marginal nature of the topics contained in the document, more than the actual novelty embedded in its content. This approach overlooks the novelty in mainstream documents and overestimates the novelty of articles in secondary or emerging fields.
A better solution to identify the novelty of an article is to consider its closest article and a neighborhood of the latter. The average distance between the article and the documents in this neighborhood is a better proxy for the innovative degree of its content because it provides a more specific context for comparisons. Specifically, we define the novelty of an article a as the average distance computed on the first 0.5th percentile of the distance (d_{ai}) distribution:

$$N_a = \sum_{i \in C_i} d_{ai},$$

where C_i is the collection of documents in the 0.5th percentile of the distribution (see Figure 1, step 3). We specifically select a very small neighborhood to detect the `local' novelty, however results are robust to the choice of a smaller (minimum of the distribution) or larger (1st percentile) neighborhood.

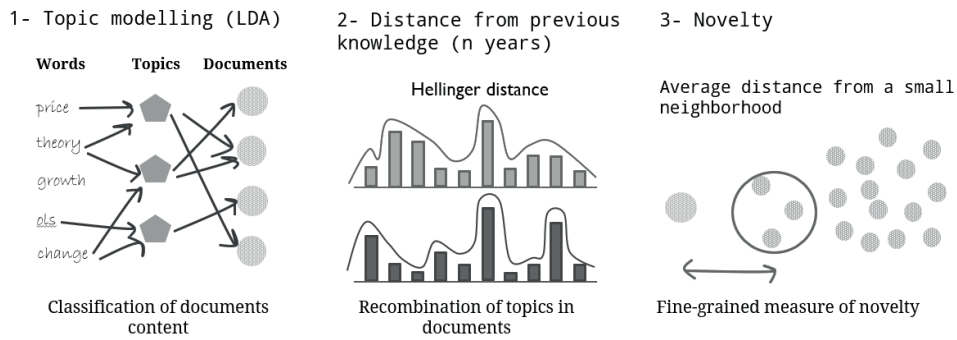


Figure 1. Steps to detect novelty in documents content

With respect to the literature on detection of breakthrough innovations, this measure has several advantages for novelty detection. Contrary to Kelly et al. (2018), it infers the latent topics in a document rather than using BoW, and focuses on a small set of articles of the previous literature to define novelty instead of computing the distance with all the existing

knowledge. Moreover, differently from Kaplan and Vakily (2015), it returns a continuous value, largely independent of the number of topics.

Acknowledgments

This template is inspired by and practically copied from the ISSI2009 template published by Birger Larssen and Jacqueline Leta.

References

Blei, D. M., A. Y. Ng, and M. I. Jordan (2003). Latent dirichlet allocation. *Journal of machine Learning research* 3, 993–1022.

Kaplan, S. and K. Vakili (2015). The double-edged sword of recombination in breakthrough innovation. *Strategic Management Journal* 36(10), 1435–1457.

Kelly, B., D. Papanikolaou, A. Seru, and M. Taddy (2018). Measuring technological innovation over the long run. *NBER working paper* 25266.

Who Plagiarizes? The Predictors of Unauthorized Borrowing in Doctoral Dissertations by Russian Scholars

Aleksandra Makeeva[1], Mikhail Sokolov[2] and Anzhelika Tsivinskaya[3]

[1] *amakeeva@eu.spb.ru*
European University at St. Petersburg, Center for Institutional Analysis of Science & Education, 6/1A Gagarinskaya Street, 191187, St. Petersburg (Russia)

[2] *msokolov@eu.spb.ru*
European University at St. Petersburg, Center for Institutional Analysis of Science & Education, 6/1A Gagarinskaya Street, 191187, St. Petersburg (Russia)

[3] *atsivinskaya@eu.spb.ru*
European University at St. Petersburg, Center for Institutional Analysis of Science & Education, 6/1A Gagarinskaya Street, 191187, St. Petersburg (Russia)

Abstract
This study reports the results of an analysis of a random sample of 2,400 Russian-language doctoral dissertations using anti-plagiarism software. We analyze which attributes (disciplinary, institutional, structural) of individuals indicate the presence of unauthorized borrowing in their dissertations. We then evaluate three general hypotheses on the causes of unauthorized borrowing – socialization (the propensity to plagiarize emerges from the lack of moral self-restriction, which correlates with other types of deviant behavior in academics), rational-choice (individuals are likely to plagiarize when the costs of abiding by the rules are prohibitively high and being caught is unlikely) and convention (inappropriate borrowing is defined by an arbitrary convention, while the spread of plagiarism is a change in convention) using our findings. Results suggest that the convention hypothesis is the most convincing as there is no evidence that high-plagiarism and low-plagiarism groups vary significantly in morals or differences in the costs versus benefits of violating rules by structural position.

Introduction
This paper studies the attributes of individuals that predict the presence of unauthorized borrowing in doctoral dissertations. We argue that the results of this analysis suggest answers to the question of why scholars plagiarize. Three hypotheses to this question are readily available for testing – *socialization, rational-choice*, and *conventional*.

The *socialization* argument is rooted in Merton's ideas on the ethos of science (Merton, 1973[1942]). According to Merton, copying other's texts verbatim is an infringement of intellectual property rights, and is therefore theft that jeopardizes institutions of science, which identify intellectual priority as essential. The theory of this framework states that plagiarizers are deviants who lack commitment to the scientific ethos and likely to break other scientific norms.

The *rational-choice* argument arises from the economic perspective on crime (Becker, 1968; on applications to scientific misconduct see Hoover, 2006; Lazetera & Zirulia, 2009). It holds that all individuals are opportunistic, but only some of them are in a position where illegal activity promises important advantages that are unavailable when adhering to norms. Individuals in such are position may be (1) academically weak and unable to produce research of an acceptable quality on their own, thus forced to accept the risk of plagiarizing to advance or else miss out on promotion opportunities or (2) in a structural position that makes detection of their misdeeds unlikely. The rational-choice hypothesis suggests that plagiarism can be found in different disciplines, including those that were the least developed during the Soviet period yet boomed after the Soviet Union's collapse due to student demand. These disciplines include the social sciences, in particular economics, management, political science and sociology. Poorly-trained personnel who rarely received quality instruction in their pedagogical position are aplenty, and there are no established communities capable of monitoring. Less unauthorized

borrowing is expected in institutions that possess a concentration of the top researchers in their field, such as the Academy of Sciences and top universities, including members of the "5/100" governmental program[1]. Those institutions enjoy stronger faculty members and likely more effective oversight. Lastly, individuals who conduct higher quality research, as measured by citation count, are presumably less likely to borrow than those whose results are less spectacular as the former can achieve prominence without risk of exposure.

The *"convention hypothesis"* states that some academics disregard the norm as they disagree with its justifications and are possibly not even aware that others support it. This perspective is rooted in constructivist theories of culture and science (see Becker, 1982 on cultural conventions in art, Gieryn, 1983 on boundary-drawing in science; Biagioli and Galison, 2014) according to which the description of certain practices as deviant is largely arbitrary. According to this perspective, textual authenticity requires any academic text to be completely original with every word written from beginning to end by the author or authors listed on the first page.[2] However, the definition "written from the first to the last word" is ambiguous. It is debated whether using certain composition (e.g. "literary review-hypotheses-methods-results-discussion and conclusion"), literary clichés ("to the best of our knowledge"), table layouts or fonts should qualify as intellectual property, and if not, why. Many authors, including several in Russia, argue that the production of substantive results is more important than the authenticity of the whole text, and that borrowing certain parts (literature reviews or methodology) is possible without diminishing the originality of the research results. In other words, the application of the textual authenticity norm requires constant evaluation between what is the "mere form" of an academic message and what is the "content". The convention hypothesis states that scholarly communities differ in how they draw this boundary. There are suppositions on which characteristics that predict plagiarism while also explaining why predictions made by the other two primary hypotheses can be falsified. The principal variable whether an individual will or will not plagiarize is that person's degree of contact with Western academia and its normative standards. Institutions conducting the best research are also likely to be more globalized (Fourcad, 2006), but this correlation is probably weaker as there are a few intervening variables.

Data and methods

To determine which hypothesis is more likely valid, we analyzed 2,400 doctoral (upper level) dissertations randomly selected from the pool of all dissertations defended in Russia between 2006 and 2016.[3] Russia is widely considered to experience major problems with academic integrity (Golunov 2014; Denisova-Schmidt, 2016). Consequently, software has been developed to identify plagiarism for Russian-language articles. We used "Antiplagiat" software, a tool that compares selected manuscripts with extensive text collections, including dissertations and academic publications stored in the Russian State Library (*Rossijskaya Gosudarstvennaja Biblioteka*). For accurate comparison, a special algorithm was used. Texts created after the defense year and papers belonging to the dissertation author were excluded. As we were only interested in borrowing within an academic social environment, we matched these dissertations with just other dissertations and research papers. We also conducted 12 qualitative interviews with individuals who have significant experience as heads or secretaries of dissertation committees.[4]

The doctoral dissertation was the unit of analysis. The dependent variable was the percentage of borrowed text and the independent variables were discipline (e.g. totaling 23 fields in which degrees in Russia are awarded, including medicine and sociology), the type of organization where the dissertation council is located, such as the Russian Academy of Science, top university (part of "5-100 program or Moscow and St. Petersburg State Universities) and scientometric indicators from individual publication profiles in the Russian Index for Scientific

Citing (percentage of publications and citations in the RISC core[5], percentage of publications and citations in foreign journals, average impact factor of journals where papers were published and cited, percentage of publications and citations in Elibrary[6] without RISC, and percentage of self-citations and citations by co-authors).
For statistical analysis of difference between different organizations one-way analysis of variance was used. Examining relation between plagiarism and scientometric indicators permutation t-test was applied. We took 10% of representatives of each discipline having the greatest and the lowest amount of borrowing and compared their publication profiles. The construction of the sample thus eliminated the influence of differences in publication profiles.

Findings

Figure 1 presents the overall distribution of plagiarism across disciplines. Two things are noticeable: first, incorrect borrowing is nearly universally present[7], and second, cases when entire texts were copy-pasted are rare. In the sample of 2,400 dissertations, we found 44 where borrowing exceeded 50%. These 44 were checked manually and three cases proved to be false positives. Overall, plagiarism exceeding 50% was identified in 41 dissertations out of 2,400 (1.7%).

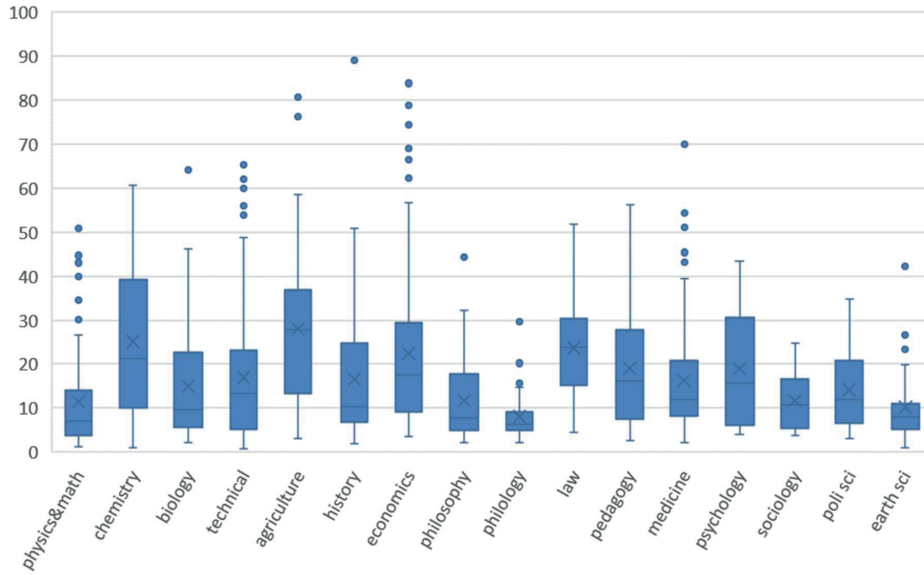


Figure 1. Distribution of plagiarism across disciplines

In contrast to the rational-choice hypothesis, there were no straightforward connections between the degree of the discipline's expansion in the market era and the level of plagiarism. Of the three disciplines with the highest levels of unauthorized borrowing, the number of defenses in chemistry has diminished since early 90-s, in agriculture it has somewhat increased, and in economics it has skyrocketed. What seems to be a more indicative factor in the case of specific disciplines is the role of disciplinary traditions, which sometimes differs in similar disciplines, e.g. chemistry and biology or economics and sociology.
Interestingly, the rational-choice logic seems applicable in another case. For relatively sizable disciplines, philology (including both literary studies and linguistics in Russia) shows the lowest share of borrowing. One might hypothesize that the reason for philologists' aversion to

plagiarism is that philology generally attracts individuals who enjoy writing and the costs of authoring the dissertation are minimal.

Table 1. Per cent of borrowing in dissertations defended at different institutions.

Organization	*Average percent plagiarized*
Top university	11.99%
Russian Academy of Sciences	15.66%
Other universities / other institutions	18.71%
ALL	**17.34%**

We found some limited support for the hypothesis predicting greater aversion to plagiarism among the faculty of research universities that are part of "5/100" program and research institutes of the Russian Academy of Sciences. A one-way analysis of variance demonstrates that there is a significant difference between Russian Academy of Science and Other (-5.036, se 0.95, p-value <0.001) and Top university and Other (-5.995, se 0.82, p-value < 0.05). However, predictions following the rational-choice hypothesis are most similar to those from the convention hypothesis as the leading institutions are the ones that, for different reasons, are the most internationalized ("5/100" universities are evaluated according to the amount of foreign students and personnel they employ, as well as the number of dual-degree programs). It is probable that leading institutions are just a gate through which international norms migrate to Russia.

We checked whether characteristics of publication profiles are somehow related with plagiarism. Self-citations and citations by co-authors metrics were included by RSCI developers, as it is known from personal communication, since they may be a product of academic activity imitation where citations are used to intentionally increase the author's and close colleagues' Hirsh-index score. However, our analysis shows that there is no difference in citations by co-authors and that norms violators have even fewer self-citations than authors from the bottom 10% plagiarism group. We argue this may be caused by norms violators who are unfamiliar with self-citation norms and thus borrow their own texts without citing unlike their less-plagiarizing colleagues.

Table 2. Differences in publication and citation performance of authors having the highest and the lowest amounts of borrowing in their work.

Variable	*Averages*		*Average treatment effect (Δ)*
	Plagiarism top 10%	*Plagiarism bottom 10%*	
Publications RISC core, %	19.96	24.99	-5.03*
Citing from RISC core, %	17.79	24.29	-6.50**
Impact factor published	0.37	0.45	-0.07*
Impact factor cited	0.42	0.52	-0.10*
Articles in foreign editions, %	3.74	6.84	-3.10***
Citing from foreign editions, %	6.94	10.4	-3.45**
Publications "Elibrary"	5.40	4.35	1.05
Citations "Elibrary"	8.35	7.70	0.65
Self-citation	20.94	28.13	-7.19***
Co-author citation	41.01	42.09	-1.08

As for Elibrary (without RISC) publications and citations, we expected a positive relation with academic norms violation as most journals excluded from the RISC are low-quality and sometimes even predatory. Nevertheless, no difference in Elibrary no-RISC publications among examined groups was revealed. This might be due to the Elibrary also incorporating conference papers and specific special editions.
While we do see some statistically significant differences in the amount of plagiarism of work by internationally published authors and those who mostly publish in second-rate domestic editions, these differences are relatively minor in absolute terms. What is more, one can suspect that those having the most impressive international publication record are also the ones most exposed to international publication norms. As a result, the disparities between them are due to the adoption of different conventions.

Conclusion

Overall, our findings cast doubt on the validity of the socialization and rational-choice hypotheses. It seems that the textual authenticity norm does not enjoy wide acceptance in Russia, and although borrowing large parts of a text (e.g. exceeding one half) is rare, recycling minor parts of other's texts is nearly universal practice (probably as much as 75% of dissertations include at least several barely re-written paragraphs from previous work). Interviews provide some detail on the rationalizations for using another's work. Some scholars justified borrowing by describing dissertation defense as a "mere formality" and decrying "senseless conventions" such as the discussion of practical applications of a dissertation on ancient history.[8] Others questioned the possibility dividing collaborative work into personalized chunks. An astronomer argued that "*an [academic] advisor and an advisee is in many senses one thing*" as a justification of the advisors' right to use their former pupils' texts in their own dissertations. There are no reasons to believe that the tendency to give and accept such explanations in any way correlates with intellectual fitness or other ethical standards.

Notes

1. A governmental program that aims to get at least five Russian universities into the top-100 of the world university rankings.
2. There is a difference between norms requiring text being fully written by its presumed author and the norm requiring the text being written for one and only one occasion. The second norm, forbidding any recycling of one's own academic text, is more restrictive than the first as it bans all forms of self-plagiarism. In this text, we deal only with the former, as using dissertation text in articles by the dissertation's author is universally regarded as permissible, if not desirable.
3. The study was conducted between March and November 2018 by the Centre for Institutional Analysis of Science and Education in collaboration with the Centre for Sociology of Education of the Russian Presidential Academy. The authors thank Katerina Guba and Nadezhda Sokolova for their contribution to this project.
4. Degrees are conferred by standing dissertation committees (*dissovet*) consisting of 25 members headed by a chair and representing the major subfields of a given discipline (e.g. sociology presently consists of "theory, history, and methodology", "social institutions and structures", "political processes", "economic processes", "culture and spiritual life", and "sociology and management").
5. The Russian Index for Scientific Citing includes a "core" of editions receiving the highest evaluations in a survey of Russian academics. It is also partially integrated with Scopus and Web of Science databases, which permits tracing publications and citations from non-Russian language editions.
6. Elibrary is a large Russian electronic library of scientific publications providing tools for search and analysis of scientific information. The library is integrated with the Russian Science

Citation Index (RISC), an instrument for measuring the publication activity of scientists and organizations.
7. Antiplagiat produces some false positives. Sometimes it counts lists of referenced literature as borrowing or does not recognize an alternative spelling of an author's name. We checked about 800 dissertations manually and received medians for disciplines of approximately five percent less. However, the manual check was rather conservative and probably underestimated the scale of borrowing. The truth thus lies somewhere between these estimates.
8. Since Soviet times, each dissertation must include a paragraph explaining its importance for the national economy. Several individuals we talked to confessed that they copy-pasted this part without any moral qualm.

References

Becker, G. (1968). Crime and Punishment: An Economic Approach. *The Journal of Political Economy,* 76(2), 169-217.

Becker, H. (1982). *Art Worlds.* University of California Press.

Biagioli, M. & Galison, P. (2014). *Scientific authorship: Credit and intellectual property in science*. Routledge.

Denisova-Schmidt, E. (2016). Corruption in Russian Higher Education. *Russian Analytical Digest*, 191, 5-9.

Fourcade, M. (2006). The Construction of a Global Profession: The Transnationalization of Economics, *American Journal of Sociology*, 112(1), 145-194.

Gieryn, T. (1983). Boundary Work and the Demarcation of Science from Non-Science: Strains and Interests in Professional Ideologies of Scientists. *American Sociological Review*, 48(6), 781-795.

Golunov, S. (2014). *The Elephant in the Room: Corruption and Cheating in Russian Universities.* Columbia University Press.

Hoover, G.A. (2006). A Game-Theoretic Model of Plagiarism, *Atlantic Economic Journal*, 34(4), 449-454.

Lacetera, N. & Zirulia, L. (2009). The economics of scientific misconduct. *The Journal of Law, Economics, & Organization*, *27*(3), 568-603.

Merton, R. (1973[1942]). The normative structure of science. In: *The Sociology of Science: Theoretical and Empirical Investigations* (268-281). Chicago: University of Chicago Press.

Data Citation and Reuse Practice in Biodiversity – Challenges of Adopting a Standard Citation Model

Nushrat Khan[1], Mike Thelwall[2] and Kayvan Kousha[3]

[1]*n.j.khan@wlv.ac.uk,* [2]*m.thelwall@wlv.ac.uk,* [3]*k.kousha@wlv.ac.uk*
University of Wolverhampton, Wulfruna St, Wolverhampton, WV1 1LY (United Kingdom)

Abstract

Openly available research data promotes reproducibility in science and results in higher citation rates for articles published with data in biological and social sciences. Even though biodiversity is one of the fields where data is frequently reused, information about how data is reused and cited is not often openly accessible from research data repositories. This study explores data citation and reuse practices in biodiversity by using openly available metadata for 43,802 datasets indexed in the Global Biodiversity Information Facility (GBIF). Quantitative analysis of dataset types and citation counts suggests that the number of studies making use of openly available biodiversity data has been increasing in a steady manner. Citation rates vary for different types of datasets based on the quality of data, and similarly to articles, it takes 2-3 years to accrue most citations for datasets. Content analysis of a random sample of unique citing articles (n=101) for 437 cited datasets in a random sample of 1000 datasets suggests that best practice for data citation is yet to be established. 26.7% of articles are mentioned the dataset in references, 12.9% are mentioned in data access statements in addition to the methods section, and only 2% are mentioned in all three sections, which is important for automatic extraction of citation information. Citation practice was inconsistent especially when a large number of subsets (12~50) were used. This calls for adoption of a standard citation model for this field to provide proper attribution when using subsets of data.

Introduction

Reproducible science is of major importance to the scientific community and the datasets reported in research articles are rich source for this. Data sharing practices seem to be more common in some fields, such as medical, forensic, and evolutionary genetics (Anagnostou, Capocasa, Milia, & Bisol, 2013). Hence, open research data initiatives have been growing quickly within different communities as data sharing aids reproducibility and reduces the chances of researchers generating data that has already been collected. However, publishing research data as first-class research output opens the door to more complex questions for researchers and policy makers –from how to define a dataset to establishing best practices of citing datasets in a specific field (Borgman, 2012; Kratz & Strasser, 2014; Starr et al., 2015; Silvello, 2018).

This study focuses on biodiversity datasets because it seems that sharing and reusing of globally collected research data are more common in this field, with primary data uses being ecological studies, taxonomic works, and phylogenetic analyses (Magurran et al., 2010; Troudet et al., 2018). For instance, a survey of 370 international researchers in biodiversity sciences indicated that most (84%) agreed that "sharing article-related data is a basic responsibility". (Huang et. al., 2012, p. 401). Global Biodiversity Information Facility (GBIF, www.gbif.org/) was used as a data source because this group has been working towards developing data publishing standards for biodiversity from an early stage (Moritz et al., 2011) and the platform holds large number of diverse datasets from different countries. Furthermore, it supports an application programming interface (API) to collect citation counts to datasets on a large scale in an automated way.

Researchers have recognized the need to provide attribution to datasets long ago. Ingwersen and Chavan (2011) suggested using Data Usage Index (DUI), an indicator based on search events and dataset download instances to demonstrate the impact of data creator and publishers. However, use of persistent identifiers for dataset was not a common practice at that time. At present, most data reuse in biodiversity uses sets of multiple datasets and these are provided

with a DOI and accession date when downloaded to cite those subsets using their in-house style. GBIF has developed a semi-automated system to assign citations to the main datasets that were included in the subsets re-used and cited by a research article.

Citing subsets complicates developing a standard model to provide the most useful information to the readers and users. As indicated by Kratz and Strasser (2014, p. 6), "…to reproduce an analysis performed on a subset of a larger dataset, the reader needs to know exactly what subset was used (e.g., a limited range of dates, only the adult subjects, wind speed but not direction). Datasets vary so widely in structure that there may not be a good general solution for describing subsets." The recently developed Dataset Search system of Google can detect the subsets mentioned indexed by DataCite. However, the original datasets need to be recognized and should be indexed by general indexing systems as well.

Citation information is not captured by most data publishing platforms due to difficulties with automating the process, caused by a lack of standards in citation styles. This makes GBIF an interesting source of information to study current data citation and reuse practices in this field and poses questions about what should be the best citation practice to make them machine-readable and how to develop a standard citation model.

Background

Citing datasets as professional reward for sharing has been mentioned by researchers in biodiversity and other fields to be a major incentive for making data openly available (Piwowar, 2011; Edmundus et al., 2012; Enke et al., 2012; Kim & Zhang, 2015; Kratz & Strasser, 2015; Sayogo & Pardo, 2013). The number of publications using GBIF data and citing GBIF has rapidly increased since 2007 (Costello et al., 2013). However, few datasets are cited in a standard format in biodiversity and the citation style is often determined by the editors for their journal (Costello et al., 2013). This is similar to life sciences data in Dryad, where the number of articles citing data in works cited section was only 8% as of 2014 (Mayo, Vision & Hull, 2016).

Previous studies have used the WoS Data Citation Index (DCI) to analyze data citation practices (Robinson-García, Jiménez-Contreras & Torres-Salinas, 2016; Park & Wolfram, 2017). However, there is evidence that DCI is relatively biased towards hard sciences and as of 2016, four repositories represented around 75% of the database (Robinson-García, Jiménez-Contreras & Torres-Salinas, 2016), although the current version of DCI indexes wider data repositories such as Figshare. Nevertheless, citation information available for each dataset on GBIF is not captured by DCI. This is an important omission, given the importance of this repository for biodiversity research and its relatively mature architecture.

The following research questions address the lack of knowledge about citation practices in GBIF – 1) Does the type of dataset or quality of information available affect citation rate? 2) How quickly do dataset citations accrue? Has the number of articles citing GBIF datasets changed over the past years? 3) Does the citation count on GBIF result from coherent citation practices? How does the use of large number of subsets impact citation practice?

Methods

This research applies an exploratory method to study the citation and reuse practice of biodiversity datasets. Quantitative analysis was used for the GBIF metadata and then content analysis was used for each unique citing article to collect information on citation location and data reuse context (Khan & Thelwall, 2019).

Data Collection

Metadata from 38,878 datasets was initially collected through the GBIF API in May 2018. The metadata fields retrieved included the dataset key, publishing organization key, dataset DOI,

dataset type, title, description, language, homepage URL, citation, citation count, creation date, and last modification date.
A random sample of 1,000 datasets was then selected for a content analysis of articles that cited datasets. About 44% (437) of datasets in the sample had at least one citing article. Between October 2018 and March 2019, a random citing article and its associated metadata was manually collected for each of the 437 datasets for full-text analysis. Download counts were also manually collected since that could not be directly retrieved through the API.
The total number of unique citing articles in the random collection was 102 as some articles appeared repeatedly for a majority of datasets. One article could not be accessed, so 101 articles were used for content analysis. The publication year, publishing journal, citation location, and contextual information of data reuse were collected for each one.
Since the data collection of citing articles was completed in 2019, an updated dataset with 43,971 datasets and a list of all citing articles for them was collected on April 6, 2019. This dataset was used to explore the distribution of all unique citing articles over publishing years.

Data Analyses

Preliminary exploration identified four types of datasets available on GBIF (GBIF, www.gbif.org/dataset-classes) – 1) Checklist datasets provide a catalogue or list of named organisms or taxa and can be used as a rapid summary or baseline inventory of taxa in a given context, 2) Occurrence datasets provide information about the location of individual organisms in time and space, 3) Sampling Event datasets contain more granular information than Occurrence datasets, often comprising of abundant information to assess community composition for broader taxonomic groups, and 4) Metadata-only datasets describe undigitized resources like those in natural history and other collections.
After de-duplicating 169 records, 43,802 datasets were used for analysis. Citation counts were analysed for all types of dataset to explore the first research question. The creation dates for each dataset were processed and average citations were calculated for the years between 2007 and 2019 for Occurrence datasets to explore how long it takes to accrue dataset citations. The list of all citing articles was de-duplicated to identify all unique articles and was used to explore the distribution over each publishing year.
A content analysis for 101 unique citing articles was conducted for a random sample of 1000 datasets for exploring the research question 4. The Spearman correlation between download and citation counts was calculated for the 437 cited datasets to help assess whether they reflect a similar type of impact.

Results

Average citations for each dataset type suggest that, Occurrence datasets are most frequently cited since they offer direct evidence of the occurrence of a species (or other taxon) at a particular place on a specified date.

Table 1. Type and number of datasets published between 2007-19 and average citations

Type	*Number*	*Percentage (%)*	*Citations per dataset*
Occurrence	16,712	93.2%	9.82
Checklist	26,216	6.4%	0.43
Metadata-only	286	0.0%	0.06
Sampling Event	588	0.4%	1.32

Prior to 2011, Occurrence datasets were the only type of datasets made available on GBIF except for two Sampling Event datasets that were published in 2007. Despite of the evidence

of a higher number of citations received by Occurrence datasets, there was a rapid increase in publishing Checklist datasets in 2016 and it is unclear why.

This study focuses on Occurrence datasets only since these are the type of datasets frequently reused and cited by articles. Figure 1 demonstrates a relatively consistent growth of Occurrence datasets. Mean citation received per occurrence dataset was 9.82, with the highest of 24.02 for occurrence datasets published in 2015 and a lowest of 0.9 for 2018. The drop in average citations per paper after 2015, indicates that, as for articles, it takes 2-3 years to accrue most citations for datasets.

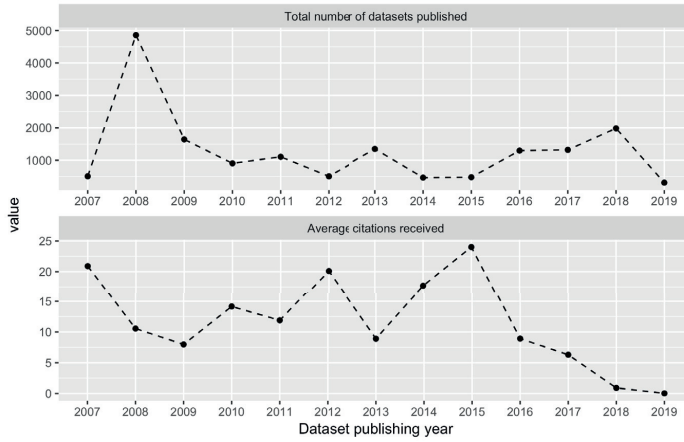


Figure 1. Number of occurrence datasets published, and average number of citations received

A correlation test was conducted for download and citation counts for the random sample of 437 cited datasets, finding a very strong positive correlation (rho = 0.787, p=0.000). Thus download counts and citation counts suggest a similar kind of impact. Because of this, early download counts might be a good indicator of longer term citation counts. Similar to the citation count findings above, Checklist datasets (n=92, average download=2610.38) had much lower download counts than Occurrence datasets (n=343, average download=5210.92) in general.

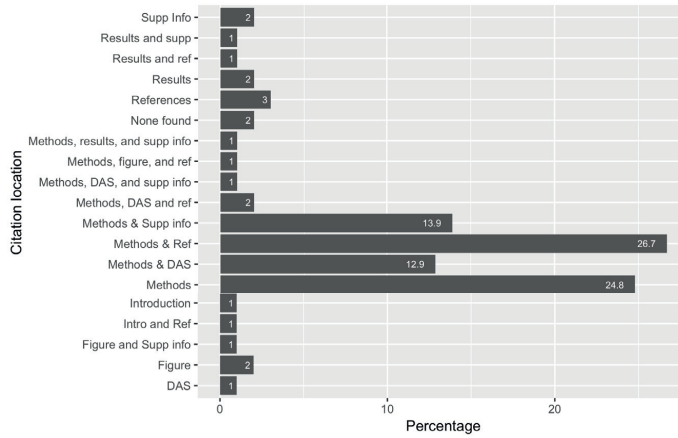


Figure 2. Citation location in randomly selected articles

A content analysis of 101 unique articles was conducted to understand citation practices in biodiversity articles citing GBIF datasets (Figure 2). Citation for GBIF dataset could not be located in two datasets. For the remaining datasets, 26.7% of the articles mentioned the dataset in their reference lists and 12.9% in data access statements in addition to the methods section, which is considered to be the standard citation practice. However, 24.8% mentioned the datasets in the methods section only within the text, which is difficult to find with indexing systems. Mentions in methods and supplementary material sections were also common (13.9%).
Most (52.5%) articles listed one subset, but some cited many (8.6% cited 50 subsets) where the number of subsets did not match for 4.9%. The non-standard citation method was used especially by articles that used large number of datasets (12~50), perhaps making it difficult to include them all in the reference section. Some articles appeared repeatedly for a majority of datasets with top 5 appearing as a citing article for more than 10,000 datasets. Those are usually the studies that used large number of records, ranging from 200 to 600 million occurrence data.

Table 2. Publication year of all citing articles mentioned on GBIF

Article Publishing Year	*Number*	*Percentage (%)*
2013	4	0.6
2014	5	0.8
2015	23	3.6
2016	70	10.9
2017	178	27.7
2018	260	40.5
2019	102	15.9

To date, 642 articles have been listed as citing article by GBIF for a total of 43,802 datasets. From the data in Table 2, it is obvious that data reuse in this field (at least from this source) has been increasing since 2013 as the number of citing articles has been growing consistently. The growth indicates the importance of openly available biodiversity data for researchers.

Discussion

This study explores data citation and reuse practice in biodiversity. It found evidence that openly available biodiversity data on GBIF is frequently reused by researchers and that the number of articles reusing and citing data retrieved from GBIF has been increasing steadily.
Citing data in references or data access statements is becoming more common but citation practices remain inconsistent across different journals. Articles using many data subsets pose extra challenges for citing in an appropriate manner. Publishing a data paper for the articles using many subsets and citing the paper itself could be a solution to this issue (Chavan & Penev, 2011). However, a refined and standard model should be adopted to address this problem when a data paper is not available. The model should also define a better way to inform the users regarding the reused subset rather than using only "GBIF Occurrence Download" as the reference title.

References

Anagnostou, P., Capocasa, M., Milia, N., & Bisol, G. D. (2013). Research data sharing: Lessons from forensic genetics. Forensic Science International: Genetics, 7(6), e117-e119.

Borgman, C. L. (2012). The conundrum of sharing research data. Journal of the American Society for Information Science and Technology, 63(6), 1059-1078.

Chavan, V., & Penev, L. (2011). The data paper: a mechanism to incentivize data publishing in biodiversity science. *BMC bioinformatics*, *12*(15), S2.

Costello, M. J., & Wieczorek, J. (2014). Best practice for biodiversity data management and publication. *Biological Conservation*, *173*, 68-73.
Costello, M. J., Michener, W. K., Gahegan, M., Zhang, Z. Q., & Bourne, P. E. (2013). Biodiversity data should be published, cited, and peer reviewed. *Trends in Ecology & Evolution*, *28*(8), 454-461.
Edmunds, S. C., Pollard, T. J., Hole, B., & Basford, A. T. (2012). Adventures in data citation: sorghum genome data exemplifies the new gold standard. *BMC research notes*, *5*(1), 223.
Enke, N., Thessen, A., Bach, K., Bendix, J., Seeger, B., & Gemeinholzer, B. (2012). The user's view on biodiversity data sharing—Investigating facts of acceptance and requirements to realize a sustainable use of research data—. *Ecological Informatics*, *11*, 25-33.
Huang, X., Hawkins, B. A., Lei, F., Miller, G. L., Favret, C., Zhang, R., & Qiao, G. (2012). Willing or unwilling to share primary biodiversity data: Results and implications of an international survey. Conservation Letters, 5(5), 399-406.
Ingwersen, P., & Chavan, V. (2011). Indicators for the Data Usage Index (DUI): an incentive for publishing primary biodiversity data through global information infrastructure. *BMC Bioinformatics*, *12*(15), S3.
Khan, N., & Thelwall, M. (2019). Dataset supporting "Data Citation and Reuse Practice in Biodiversity". figshare. Dataset. 10.6084/m9.figshare.8181098.v1
Kim, Y., & Zhang, P. (2015). Understanding data sharing behaviors of STEM researchers: The roles of attitudes, norms, and data repositories. *Library & Information Science Research*, *37*(3), 189-200.
Kratz, J., & Strasser, C. (2014). Data publication consensus and controversies. *F1000Research*, *3*.
Kratz, J. E., & Strasser, C. (2015). Making data count. *Scientific data*, *2*.
Kratz, J. E., & Strasser, C. (2015). Researcher perspectives on publication and peer review of data. *PLoS One*, *10*(2), e0117619.
Magurran, A. E., Baillie, S. R., Buckland, S. T., Dick, J. M., Elston, D. A., Scott, E. M., Smith, R. I., Somerfield, P. J., & Watt, A. D. (2010). Long-term datasets in biodiversity research and monitoring: assessing change in ecological communities through time. *Trends in ecology & evolution*, *25*(10), 574-582.
Mayo, C., Vision, T. J., & Hull, E. A. (2016). The location of the citation: changing practices in how publications cite original data in the Dryad Digital Repository. *International Journal of Digital Curation*, *11*(1), 150-155.
Moritz, T., Krishnan, S., Roberts, D., Ingwersen, P., Agosti, D., Penev, L., Cockerill, M., & Chavan, V. (2011). Towards mainstreaming of biodiversity data publishing: recommendations of the GBIF Data Publishing Framework Task Group. *BMC Bioinformatics*, *12*(15), S1.
Park, H., & Wolfram, D. (2017). An examination of research data sharing and re-use: implications for data citation practice. *Scientometrics*, *111*(1), 443-461.
Piwowar, H. A. (2011). Who shares? Who doesn't? Factors associated with openly archiving raw research data. *PloS one*, *6*(7), e18657.
Piwowar, H. A., & Vision, T. J. (2013). Data reuse and the open data citation advantage. *PeerJ*, *1*, e175.
Sayogo, D. S., & Pardo, T. A. (2013). Exploring the determinants of scientific data sharing: Understanding the motivation to publish research data. *Government Information Quarterly*, *30*, S19-S31.
Silvello, G. (2018). Theory and practice of data citation. *Journal of the Association for Information Science and Technology*, *69*(1), 6-20.
Starr, J., Castro, E., Crosas, M., Dumontier, M., Downs, R., Duerr, R., Haak, L., Haendel, M., Herman, I., Hodson, S., Hourclé, J., Kratz, J., Lin, J., Nielsen, L., Nurnberger, A., Proell, S., Rauber, A., Sacchi, S., Smith, A., Taylor, M. and Clark, T. (2015). Achieving human and machine accessibility of cited data in scholarly publications. *PeerJ Computer Science*, 1, p.e1.
Robinson-García, N., Jiménez-Contreras, E., & Torres-Salinas, D. (2016). Analyzing data citation practices using the data citation index. *Journal of the Association for Information Science and Technology*, *67*(12), 2964-2975.
Troudet, J., Vignes-Lebbe, R., Grandcolas, P., & Legendre, F. (2018). The increasing disconnection of primary biodiversity data from specimens: How does it happen and how to handle it?. *Systematic biology*.

A comparison of three individual multidisciplinarity indices based on the diversity of the Scopus subject areas, of the bibliography and of the citing papers of an author's documents

Ugo Moschini[1], Elena Fenialdi[2], Cinzia Daraio[3], Giancarlo Ruocco[1] and Elisa Molinari[1]

[1] *[ugo.moschini, elisa.molinari, giancarlo.ruocco]@iit.it*
Istituto Italiano di Tecnologia, Genoa, (Italy)

[2] *elena.fenialdi@gmail.com*
University of Genoa, Genoa (Italy)

[3] *daraio@dis.uniroma1.it*
Sapienza University of Rome, Rome (Italy)

Abstract

In this paper, we compare the distribution of Elsevier Scopus subject areas related to three cases: the authors' documents, their bibliographical references and their citing documents. We compute the complement of the Hirschman-Herfindahl (HH) index, as a measure of diversity, for each case. We analyse an overall sample of 120 researchers belonging to two groups, one from the Italian Institute of Technology (whose work is expected to be highly multidisciplinary) and one from the National Institute for Nuclear Physics (whose work is expected to be much less multidisciplinary). We show that the two groups are distinguishable through the measured diversity index values. In particular, by using the subject areas of authors' bibliographical references we obtain a better identification of the two groups than relying on the subject areas of the author's documents. This result seems interesting for assessing the interdisciplinarity of younger researchers with scarce scientific output and few citations.

Introduction

Nowadays, measuring and understanding *multidisciplinary* research is of prominent interest for both researchers and funders or evaluators (Wagner et al., 2011). Multidisciplinary works and consequent applications are generally considered to have a stronger impact on the society and the scientific development (Rafols & Meyer, 2010). In spite of this growing interest, both defining and measuring multidisciplinarity is not trivial. Several definitions have been given to describe the multiple modalities in which disciplines can interact (Stokols et al. 2003; Choi & Pak, 2006; Porter et al., 2007). We use the term multidisciplinarity in a broad sense, indicating that elements from different disciplines are present. The term "multidisciplinarity" is used in the following and the concept of interdisciplinarity is inherent to it.

Since many years, a majority of quantitative indicators of the degree of multidisciplinarity of a researcher are based on bibliometric methods (Porter et al., 2007). These methods are commonly classified in bottom-up and top-down approaches (Wagner et al., 2011). Bottom-up approaches are based on grouping and forming sets of articles according to a criterion, like building a co-citation network (Boyack & Klavans, 2010). This approach is suitable for finding emerging fields, in which there is no classification available, very few publications and no a priori taxonomy (Leydesdorff, 2007; Leydesdorff et al., 2013). Top-down approaches are dependent on some kind of classification available. They are suitable for a large-scale analysis, especially when dealing with big amount of data (Porter & Rafols, 2009; Leydesdorff et al., 2013). In fact, published manuscripts are commonly included and indexed in several databases such as Scopus and Web of Science. Elsevier's Scopus is one of the most important databases due to its large coverage catalogue. A feature offered by such databases is that main topics or research areas are assigned to articles and journals via (semi-)automatic methods. The topics are named *subject areas (SA)* in Elsevier's Scopus database. We use the top-down approach

based on the Scopus catalogue in this paper. Bibliometric data and their classification into topics are the raw material used to measure interdisciplinarity.

In 2007, Stirling introduced a general framework for analysing diversity in science, that takes into account variety, balance and disparity (Stirling, 2007). His idea is to take into account not only the number of disciplines in articles (or in their reference lists), but also the *distance* between them: he suggested to apply the Rao index developed in biology (Rao, 1982) to measure research multidisciplinarity, defining the so-called Rao-Stirling index. The Rao-Stirling index is a popular indicator of interdisciplinarity, used in many works (Rafols & Meyer, 2010; Leydesdorff & Rafols, 2011, Leydesdorff et al., 2013; Wagner et al., 2011). However, creating a meaningful distance measure is far from being an easy task (Rafols & Meyer, 2010). This requires the selection of a context (a set of papers), the identification of attributes on which a distance measure is based and so on.

In our work, we are going to analyse the behaviour of the Rao-Stirling index when no distance measure is used. This index was demonstrated to be equal to the Hirschman- Herfindahl (HH) index, very popular in economics (Rhoades, 1993; Rousseau, 2018), with different choices of parameters (Porter & Rafols, 2009). We use the HH index, also known as Simpson (Simpson, 1949) index, to analyse the differences in the distribution of subject areas in three cases: the subject areas of an author's documents, of his/her documents' bibliographical references and the citing documents of his/her documents. The evaluation is conducted on two sets of researchers and their publications in the years 2010-2018. The first set is made of researchers working at the Italian Institute of Technology (IIT) that is a Foundation established by Law no. 326/2003, and financed by the Italian State to conduct scientific research in the public interest, for the purpose of technological development (see further information at: https://www.iit.it/about-us/institute) and according to its strategic plan is engaged in highly multidisciplinary research activities. These researchers, then, are expected to be highly multidisciplinary. The second set contains researchers affiliated with the National Institute for Nuclear Physics (INFN, further information are available at http://home.infn.it/en/) in Italy. These researchers are expected to be much less multidisciplinary.

Distribution of subject areas

For the two chosen sets of researchers, we considered only the documents published in the period 2010-2018, as indexed in Scopus. Documents were accessed through the Author Scopus identifiers of each researcher profile on Scopus: the subject areas were retrieved via a script based on the Scopus Application Programming Interface (Elsevier Developers, 2018). The Scopus subject areas (SAs) are classified in 334 categories. Scopus assigns a varying number of subject areas to the *source medium* in which a document appears (i.e., journal, conference proceedings, book, ...), rather than to the document itself. Since Scopus assigns subject areas to journals or conferences and not to the article itself, the assumption is that an article inherits the SAs of the source where it is published. In our analysis, we distinguished three different ways to consider the subject areas that could potentially describe the disciplines belonging to the scientific production of a researcher:

1. SAs of the documents written by a researcher (**ARG**).
2. SAs of the bibliographical references of every document written by a researcher (**BIB**).
3. SAs of the publications that cite the documents written by a researcher (**CIT**).

We computed the percentage of occurrence of every subject area in each case. A researcher is then described by three vectors of 334 components, each component representing a given SA. The category MULT indicating journals that publish articles coming from several disciplines, rather than inherently multidisciplinary articles, was removed from the dataset.

The diversity index used in this paper

To measure the diversity of authors' research outputs we used the complement of the HH index, also known as Simpson diversity index. The HH index is a very popular concentration index in economics used to measure how the market share is distributed among companies. In our context, the market share distribution among companies translates into the subject areas related to a researcher, more specifically into the vector of percentages.

Let us call V a vector containing percentages of SAs and let us assume its values are normalised to range between 0 and 1. We are interested in the complement of the HH as a measure of *diversity* HH index. In fact, it is easy to see that less concentration means more multidisciplinarity. Let us define $CHH(V)$, the index computed on a vector V, as:

$$CHH(V) = 1 - \sum_{i=1}^{N} v_i^2 \qquad , v_i \in V$$

The value returned by the equation above takes values in the range 0-$1/N$.

Experiments and discussion

We choose two separate sets of researchers for our experiments. The first set contains 64 researchers affiliated at the Italian Institute of Technology (IIT), expected to show an high multidisciplinarity degree: their scientific output often combines, for example, robotics with life sciences, medicinal chemistry with biology, and so on. The second set contains 56 researchers affiliated at the National Institute for Nuclear Physics (INFN), whose work has inherently a narrower scope: hence, less interdisciplinarity is expected. They are senior researchers, i.e. with a number of publications at least larger than 20. We are interested in verifying if the complement of the HH, our diversity measure, shows higher values for the IIT set compared to the INFN set, as should follow from our a priori selection. The aim of our tests is to check the behaviour of this index, i.e. its discriminatory power between the two groups of researchers, when considering i) the subject areas of the author's documents, ii) or the bibliographical references of author's documents , iii) or the citing documents of author's documents. We then computed the complement of the HH index on the vectors of percentages of Scopus subject areas, for the 120 researchers, for the three cases ARG, BIB and CIT. Only documents published within the years 2010-2018 were considered, to avoid a potential correlation between multidisciplinarity and the age of a researcher. Each of the 120 researchers published documents before 2010: these documents were simply discarded from the analysis.

Figure 1 shows the results obtained after computing the complement of the HH index, multiplied by 100 for convenience. In the plots, the IIT researchers are indicated in *black* colour, while the INFN researchers in *grey*. The left plot of Figure 1 shows the index values unsorted and grouped by researchers' affiliation, for the three cases. In the plots, we notice that the range of index values for INFN researchers is much broader in the BIB and CIT case than in the ARG case. The mean value for the INFN researchers in the ARG case is much more different from the BIB and CIT case. Instead, the mean values for IIT researchers fall in a similar range in each of the three cases. Table 1 shows in details the range of indexes and the mean/median values for the IIT and INFN researchers for the three cases. In the right plot of Figure 1, the index values are shown sorted, for each case. Ideally, all the IIT researchers would stay above the INFN researchers, due to the fact that they are more multidisciplinary. We notice that the two sets are split up to a good extent. While an ideal situation occurs in the CIT case, unfortunately it does not hold in ARG and BIB: there are in total three researchers who cross the boundary of their own group, *overlapping* with the others. Interestingly, the very same researcher is common to both ARG and BIB. The actual scientific production of these "overlapping" authors should be further investigated: their research could be not so monothematic or multidisciplinary as we thought a priori.The limit between what can be

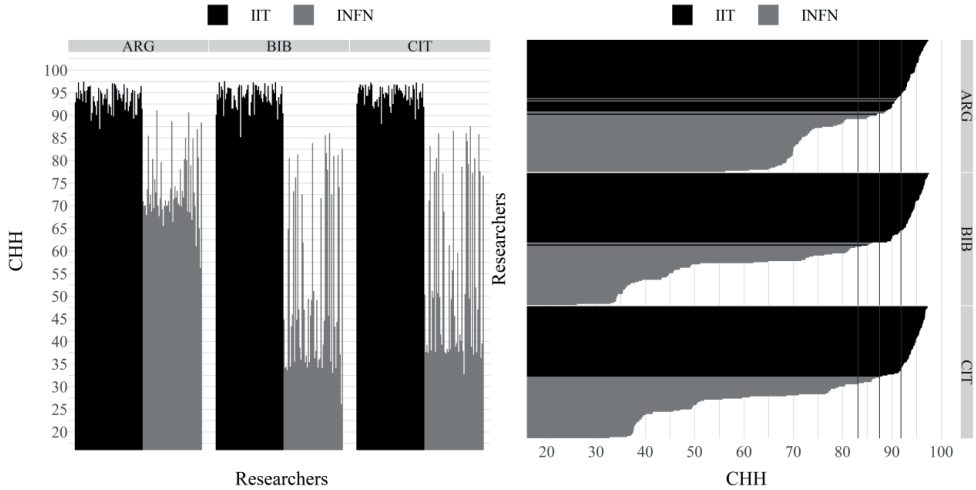


Figure 1. (Left) values of the complement of the HH (CHH) index for ARG, BIB and CIT on the 120 researchers; (Right) sorted values of the complement of the HH (CHH) index on the 120 researchers.

Case	*Researchers*	*Min*	*Max*	*Median*	*Mean*
ARG	IIT	0.8697	0.9742	0.9378	0.9371
ARG	*INFN*	*0.5622*	*0.9113*	*0.7132*	*0.7362*
BIB	IIT	0.8513	0.9756	0.9457	0.9413
BIB	*INFN*	*0.2611*	*0.8605*	*0.4480*	*0.5165*
CIT	IIT	0.8807	0.9727	0.9485	0.9452
CIT	*INFN*	*0.3276*	*0.8760*	*0.4957*	*0.5508*

Table 1. Min/max, median and mean values of the complement of the HH index for ARG, BIB and CIT and the IIT/INFN researchers.

multidisciplinary and what cannot be is unlikely a rigid and exact boundary: it is a promising sign, however, that researchers of a set overlap with researchers of the other set only near the boundary region If we select a few researchers less multidisciplinary and less "monodisciplinary" (i.e., at the boundary of the black and grey "regions" of the left plot in Figure 1), we find that the sets made of these names have names in common: such researchers have approximately the same ranking position in the three cases. We extended our analysis to study how the relative ranking positions of the researchers change throughout ARG, BIB and CIT. We measured that 45.83% of the researchers remained in a range of ±5 positions, comparing the ranking of ARG vs BIB. The percentage goes up to 60.83% for BIB vs CIT. If the range considered is ±10 positions, the percentage is 62.5% and 80.83%, considering ARG vs BIB and BIB vs CIT. Also, the experiments showed that the ranking changes less on the INFN set, although the size of the sets is not enough for drawing precise conclusions. Overall, only about 10% of the researchers maintain the same position in the ranking across the cases. By and large, it appears that researchers keep roughly their position (about 85% of the researchers stay in a range of ±15 positions) in the three cases. We computed also the Kendall rank correlation among the vectors of percentages in the three cases. On the whole set of 120 researchers, Kendall's correlation is 0.78 and 0.81, for ARG vs BIB and ARG vs CIT, respectively. Considering the IIT set, the correlation decreases to 0.61 and 0.67, while on the INFN set it goes down to 0.54 and 0.58. This means that the order of the researchers generated

by the index based on the three different vectors (ARG, BIB, CIT) varies largely within each set, but the overall order keeps the separation of the two sets, hence we measure higher correlation when considering all the researchers together. This happens because Kendall's correlation counts how many concordant or discordant couples are present: an increase in correlation means that the percentage of concordant couples (couples of researchers for whom the order does not change) is higher over all of the researchers.

We observe that ARG, BIB and CIT may measure different kinds of multidisciplinarity: the citations indicate the visibility of an author's work among various disciplines rather than the multiple disciplines a work is based on. Integration of different disciplines is probably better shown by the references cited by an article in its bibliography (Porter et al., 2007): an authors' own publications would instead indicate how diverse the individual production is. From our study, the use of bibliographical references (the BIB case) showed a better separation between IIT and INFN than the ARG case: the mean/median are further apart and less researchers end up in the wrong region. The use of the bibliographical references gives good clues about multidisciplinarity and also allows for measuring younger researchers whose scientific output is still limited or who are not cited many times: the bibliography is normally composed by a large number of entries, making it suitable in these situations, too. It would also be interesting to check if a correlation exists between long reference lists and a high multidisciplinarity of the CIT case.

Conclusions and future work

In literature, the complement of the HH index, also known as Simpson diversity (see e.g. Rousseau, 2018) has been often used to measure multidisciplinarity, although, to the best of our knowledge, not on Scopus subject areas (SAs). In our work, we analysed its performance on data related to IIT and INFN researchers. The two sets were chosen so that the former would potentially show a more multidisciplinary scientific production, while the latter less multidisciplinarity. We considered three cases of subject areas (SAs) from Scopus: the SAs of the documents written by an author, those of the bibliographical references of an author's documents and those of the citing documents of an author's documents. The results are promising: the he two sets are differentiated to a good extent by the complement of the HH index. Analysing the SAs of the bibliography of the publications of a researcher seems the best way forward. In this case, the distinction between the two groups of researchers is higher than the one obtained using the SAs of the researchers' articles. Moreover, the bigger size of bibliographical references would make the method suitable for researchers with a low number of publications and few citations. While the distinction among the two sets is kept, the value of the diversity index (and its relative ranking) of a researcher changes in each case: the three cases might measure different aspects of multidisciplinarity.

More definitive conclusions could be drawn extending the analysis and using statistical tests over a larger sample of researchers, working in international institutes and diverse scientific fields: that is currently a subject of ongoing research in our group. Analysing the researchers in a specific research area could give important insights on the expected range of the multidisciplinarity index relevant to a specific area. In the Italian system, academic research areas translate into officially pre-assigned Scientific Disciplinary Sectors (Settore Scientifico Disciplinare, SSD). This allows to analyse multidisciplinarity considering the official field classification of authors (Abramo et al. 2012). Nonetheless, an official disciplinary classification for other public research centers, such as the IIT, does not exist. For this reason, comparing the multidisciplinarity using not only the official classification of authors (Abramo et al. 2018) but also other approaches, such as the one used in this paper (namely diversity of the Scopus subject areas of documents, of the bibliography and of the citing papers of an author's documents), may be useful.

The comparison we have carried out in this paper could be further extended to check the robustness of the official disciplinary classification of the Italian academic system and may provide a more general classification able to embrace also researchers from institutions for which the official disciplinary classification is not available. Moreover, it would be interesting to explore how the multidisciplinarity measures change within a specific SSD and discover possible patterns of multidisciplinarity relative to each scientific disciplinary sector.

References

Abramo, G., D'Angelo, C.A., & Di Costa, F. (2012). Identifying interdisciplinarity through the disciplinary classification of coauthors of scientific publications. Journal of the American Society for Information Science and Technology, 63(11), 2206–2222.

Abramo, G., D'Angelo, C. A., & Zhang, L. (2018). A comparison of two approaches for measuring interdisciplinary research output: The disciplinary diversity of authors vs the disciplinary diversity of the reference list. Journal of Informetrics, 12(4), 1182-1193.

Boyack, K. W., & Klavans, R. (2010). Co-citation analysis, bibliographic coupling, and direct citation: Which citation approach represents the research front most accurately? Journal of the American Society for Information Science and Technology, 61 (12), 2389-2404.

Choi, B. C., & Pak, A. W. (2006). Multidisciplinarity, interdisciplinarity and transdisciplinarity in health research, services, education and policy: 1. Definitions, objectives, and evidence of effectiveness. Clinical and Investigative Medicine. Medecine Clinique et Experimentale, 29(6), 351–364.

Elsevier Developers. (2018). https://dev.elsevier.com/. ([Online; accessed 04-July-2018])

Leydesdorff, L. (2007). Betweenness centrality as an indicator of the interdisciplinarity of scientific journals. Journal of the American Society for Information Science and Technology, 58(9), 1303-1319.

Leydesdorff, L., & Rafols, I. (2011). Indicators of the interdisciplinarity of journals: Diversity, centrality, and citations. Journal of Informetrics, 5 (1), 87 - 100.

Leydesdorff, L., Rafols I., & Chen, C. (2013). Interactive overlays of journals and the measurement of interdisciplinarity on the basis of aggregated journal - journal citations. Journal of the American Society for In- formation Science and Technology, 64 (12), 2573-2586.

Porter, A. L., Cohen, A. S., David Roessner, J., & Perreault, M. (2007). Measuring researcher interdisciplinarity. Scientometrics, 72 (1), 117-147.

Porter, A. L., & Rafols, I. (2009). Is science becoming more interdisciplinary? Measuring and mapping six research fields over time. Scientometrics, 81 (3), 719.

Rafols, I., & Meyer, M. (2010). Diversity measures and network centralities as indicators of interdisciplinarity: Case studies in bionanoscience. Scientometrics 82 (2), 263-287.

Rao, C. R. (1982). Diversity: Its measurement, decomposition, apportionment and analysis. Sankhy: The Indian Journal of Statistics, Series A (1961-2002), 44 (1), 1-22.

Rhoades, S. A. (1993). The Herfindahl-Hirschman index. Federal Reserve Bulletin(Mar), 188-189.

Rousseau, R. (2018). The repeat rate: from Hirschman to Stirling. Scientometrics, vol. 116(1), 645-653.

Simpson, E. H. (1949). Measurement of diversity. Nature, 163(4148), 688.

Stirling, A. (2007). A general framework for analysing diversity in science, technology and society. Journal of The Royal Society Interface, 4 (15), 707-719.

Stokols, D., Fuqua, J., Gress, J. Harvey, R., Phillips, K., Baezconde-Garbanati, L., ... Trochim, W. (2003). Evaluating transdisciplinary science. Nicotine & Tobacco Research 5(6), 21 – 39.

Wagner, C. S., Roessner, J. D., Bobb, K., Klein, J. T., Boyack, K. W., Keyton, J., ... Borner, K. (2011). Approaches to understanding and measuring interdisciplinary scientific research (IDR): A review of the literature. Journal of Informetrics, 5 (1), 14-26.

Mapping an emerging research subject: case of microbiota concept

Abdelghani Maddi[1], David Sapinho[2], Lesya Baudoin[3]

[1] abdelghani.maddi@hceres.fr
Observatoire des Sciences et Techniques, Hcéres, Rue Albert Einstein, Paris, 75013 (France); CEPN, UMR-CNRS 723, Université Paris 13

[2] david.sapinho@hceres.fr
Observatoire des Sciences et Techniques, Hcéres, Rue Albert Einstein, Paris, 75013 (France)

[3] lesya.baudoin@hceres.fr
Observatoire des Sciences et Techniques, Hcéres, Rue Albert Einstein, Paris, 75013 (France)

Abstract

Microbiota research has experienced an exponential growth over the last 10 years. Globally, the number of publications dealing with microbiota has increased by more than 40 between 1999 and 2017. This expansion is the corollary of several technological advances supported by public and private initiatives to finance microbiota projects. Our study combine qualitative and quantitative approaches, on the one hand, to map the topical landscape of the microbiota-related publications and its dynamics, and on the other hand, to analyse the capacity of countries to integrate a new research field through the increase of scientific publications and their scientific specialization by topic. Using the Web of Science database, 28489 articles and reviews published between 1999 and 2017 and containing the word “microbiota” in title, abstract or keywords were retrieved. A text mining analysis (topic modelling) of the "keywords" fields allowed us to identify 23 topics with contrasting dynamics over the period, suggesting that the microbiota research has entered the translational phase. Publication trends by country in microbiota research show specific patterns, different from those observed for all disciplines. Country specialization analysis reveals three groups: clinical, applied and weakly specialized.

Introduction

Recent discoveries highlighting the major impact of gut bacteria on various health conditions have made microbiota a subject of rising interest both to the scientific community and to the general public. Given its potential as a new therapeutic target, a number of research efforts have been funded by national and international programs. The term “microbiota” has become somewhat of a buzzword in today’s science news. Mainstream media headlines mention microbiota as a “new concept” or even as a “new organ” whereas the food industry uses it as a marketing argument. However, the term “microbiota” can be found in scientific publications since the beginning of the 20th century, although its scope evolved. The term’s use remained quite rare until the mid-1990s. It is over the past two decades that the number of publications has undergone the most remarkable increase.

Growing number of papers on particular subject reflects a rise of research activity and could be considered as an indicator of emerging area of research (Guo, Weingart, & Börner, 2011).

In this context, the capacity of countries to seize the novelty and to integrate a nascent research topic has important implications for research policy.

Despite obvious research interest, the field of microbiota has rarely been approached from the bibliometric perspective. Although several studies have been conducted, they are limited to specific aspects as obesity, diabetes or wastewater treatment (Garrido-Cardenas, Polo-López, & Oller-Alberola, 2017; Tian, Li, Lian, & Tong, 2017; Yao et al., 2018) but the field as a whole has not yet been analysed.

The aims of our study are: (1) to sketch out a general landscape of research dealing with microbiota using topic modelling and to outline its evolution; (2) to analyse the contribution of different countries into this research fields through the publications and to provide insights into their respective topic specialization. The time frame of analysis has been restricted to the period 1999-2017, corresponding both to the strongest increase and to the broadest diversification of the use of the term.

Data and Methods

Source of Data

The data has been extracted from *Observatoire des sciences et Techniques' (OST)* in-house database. OST database includes five indexes of the Web of Science (WoS) available from Clarivate Analytics: Science Citations Index Extended (SCIE), Social Sciences Citations Index (SSCI), Arts & Humanities Citations Index (AHCI), Conference Proceedings Citation Index-Social Science & Humanities (CPCI-SSH) and Conference Proceedings Citation Index-Science (CPCI-S). The data used correspond to the WoS content indexed through the week of March 26, 2018 and the completeness for 2017 is estimated as 95% on average.

The search query used for data extraction contains the word "microbiota" in the title, abstract or keywords. To avoid the few false positive results (11 documents) due to the homonymous term "Microbiota decussata" (Latin name of a coniferous plant), the documents containing word "decussata" were filtered out. Only publications indexed as "article" or "review" have been taken into consideration. The query returned 28,489 publications for the period 1999 - 2017.

Topic modeling

Topic modeling is one of the widespread text-mining techniques used in scientometric studies. Most often, topic models rely on frequency lists of unigrams extracted from unstructured textual fields as title, abstract or full text documents. However, not all words in a text have equal semantic weight, and unigrams can be ambiguous. We assumed that the information contained in keywords fields capture the main themes of the paper in a more accurate and concise way. Besides, keywords often consist of multi-word terms that are easier to interpret and semantically richer. Web of Science records contain two keywords fields: Author Keywords provided by authors and Keywords Plus automatically extracted from titles of the

cited documents by Clarivate Analytics. Both are uncontrolled keywords, which means that no normalization has been applied to these vocabularies. Although keywords are generated by a computer algorithm, the study of Zhang et al. (Zhang et al., 2016) has demonstrated that Keywords Plus are as effective as Author Keywords for discovering the semantic structure of scientific literature. Furthermore, taking both keywords fields allowed us to broaden the document coverage, since some papers do not contain Authors Keywords and others do not contain Keyword Plus.

For this analysis, terms were extracted as strings delimited by semicolons from Keywords Plus and Author Keywords fields from the overall set of articles (N = 28,489), of which 89 were excluded since they did not contain any keywords field.

A 5-step data preprocessing was applied to the initial list of keywords (n = 60,818), in order to reduce the dimensionality of the document-term matrix by conflating the candidate terms (Figure 1):

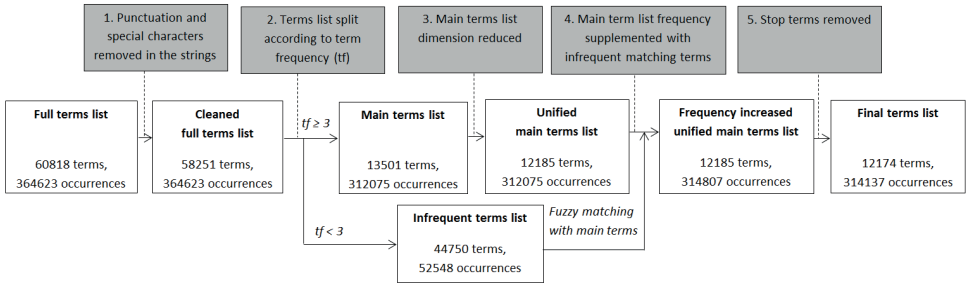


Figure 1: process of building the list of keywords

1. The punctuation and special characters were removed from the initial strings. In contrast with widespread use, the numbers were not removed since they may be often semantic carriers (e.g. “E. coli” is generally harmless while the strain “E. coli O157” is highly pathogenic).
2. The resulting list was divided into two subsets, main list containing the terms occurring 3 or more times and secondary list with less frequent terms (<3).
3. A pairwise comparison of each couple of terms from the main list was made to reduce the number of terms by conflating singular and plural forms and orthographical or syntactic variants.
4. The infrequent terms list was harvested to detect candidates for conflating with the main terms, increasing their frequency.
5. Finally, stop words were removed.

To summarize the semantic information extracted from the corpus, the Latent Dirichlet Allocation (LDA) method was applied (Blei, Ng, & Jordan, 2003). This unsupervised

algorithm allows the detection of latent thematic structures (topics) from a set of selected terms in documents (Blei, 2012; Blei et al., 2003). The method assumes that a probabilistic distribution determines how terms, considered as realizations of latent variables, are randomly attributed to documents. Under this assumption, the observed document-term matrix must be seen as the product of both estimated document-topic and topic-term matrices, where the first gives the proportion of topics within each document, while the second gives the proportion of terms in each topic. Each document can thus be associated with several topics. The weight of a topic is defined as a sum of weights of this topic in the documents.

The LDA model was fitted with the Mallet package (McCallum, 2002). R packages "LDAvis" and "topicmodels" packages were used to analyse data, display the results and provide an interactive visualization of topics (Sievert & Shirley, 2014). The proximity in the figure indicates similarity between terms distribution of topics.

The number of topics retained for the analysis was chosen with regard to the perplexity of models provided with different number of topics, a lower perplexity score indicating better generalization performance (Blei et al., 2003). The perplexity decreases when the number of topics increases. However, in a certain point, the interpretability of topics becomes complex. In our case, the best compromise was between 20 and 25 topics. The six models were tested, and the model with 23 topics was selected as it was the most relevant to interpret. In accordance with the conclusions of Chang et al (2009) we consider that the human interpretability is as important as metrics in assessing coherence and relevance of defining the number of topics; therefore both methods were combined.

The evolution of the weight of topics per year is represented in an alluvial graph. These types of graph allow showing both the evolution of the volume as well as the ranks of the different topics by year.

Country analysis

We used the topic-documents matrix to calculate the fractional number of publications per country. Thus, for each document, the weight of the topic is multiplied by the weight of the country. For example, for an article co-authored by three authors, one French and two Dutch, and allocated for 90% to topic 21 and 10% to topic 10, the publication will be counted as follows:

	Topic weight	
Country weight	**Topic 10 : 0.1**	**Topic 21 : 0.9**
France : 1/3 = 0.33	0.33*0.1 = 0.033	0.33*0.9 = 0.297
The Netherlands : 2/3 = 0.66	0.66*0.1 = 0.066	0.66*0.9 = 0.594

These specific weights were then summed to estimate contributions per topic, per year and per country, and to infer the specialisation by topic for the top 20 producing countries. Scientific specialisation index describes the relative research intensity of the country in the topic with respect to the world. It is defined by the share of the topic in the country's publications

(fractional counting), divided by the share of the same topic in the publications worldwide. It is equal to 1 if the country's share in the topic equals the country's share in all fields indicating no specialisation. The higher the specialisation index is above 1, the more specialised the country is in that topic.

Results

Trends in scientific publications in microbiota research

The preliminary search conducted in three online major bibliographic databases – Web of Science, Scopus and PubMed (including all document types) – revealed a similar growth rate of the microbiota-referring publications, suggesting that they adequately reflect the literature on the subject and are equally fit for analysis. Interestingly, all three databases' first mention of microbiota refers to the same 1956 paper (Lackey, 1956). From that point on until the mid-1970s, less than 10 papers dealing with microbiota appeared annually, and it took another 20 years before reaching 100 papers per year. Our study using the Web of Science data covers the last 18 years. Figure 2 shows that, between 1999 and 2017, the general growth trend has been exponential.

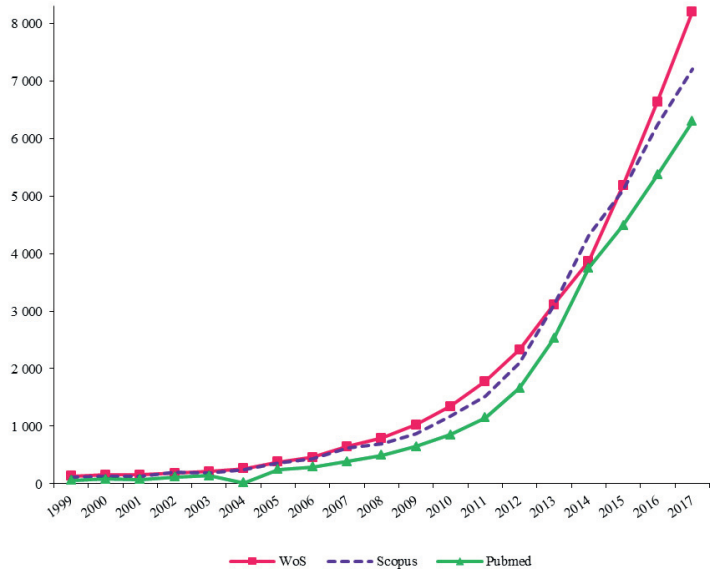


Figure 2: General growth trend of "microbiota" publications

We analysed the contribution of different countries into the microbiota research. Overall, 151 countries have contributed to this field during the studied period. Along with the increase of the publication output, the number of countries involved has grown from year to year, from 28 distinct countries in 1999 till 119 in 2017.

Figure 3 shows the output of the microbiota-related papers for the top fifteen producing countries in the field. For readability, the timeline was truncated, since the publication number in 1999-2008 was too low comparatively to the following years. Despite the steady increase in annual publications the observed country-level trends have changed significantly over the past ten years. Apart from the United States, which ranks first throughout the period, there were significant changes in volume of publications per country. China's output has progressed; since 2013 it became the second producing country in this field, while it was ranked 11th in 2009. Italy and France are positioned among the major players in microbiota research. They have a very similar number of publications in 2017 and rank respectively, 3rd and 4th. By comparison, taking into account the overall publications output, they come behind United Kingdom (8th in microbiota field), Germany, Japan and India (Observatoire des Sciences et Techniques, 2018).

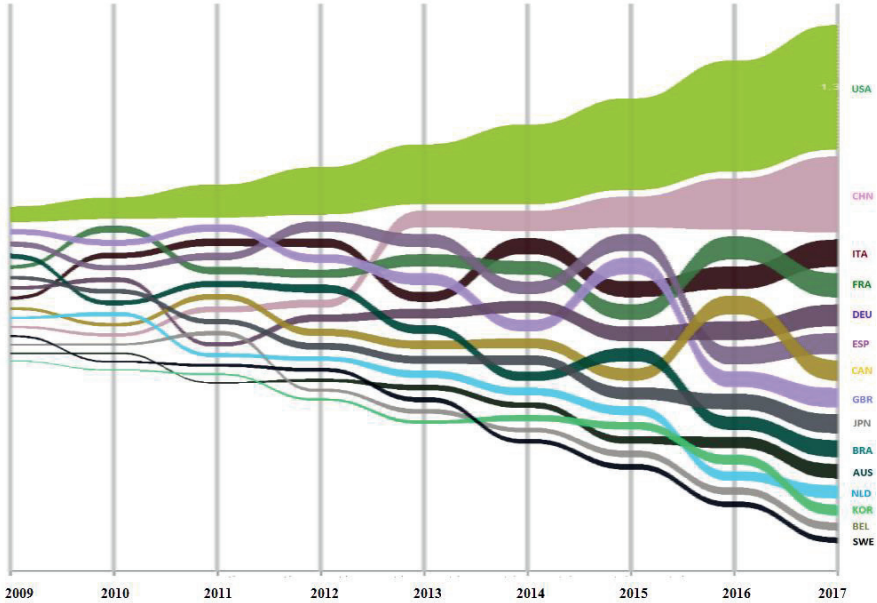


Figure 3: trends in scientific publications in microbiota research, top 15 countries

Topic analysis

The document-level analysis allowed us to investigate the content of the documents set through their vocabularies and to produce a summary map of the field. The model with 23 topics was found to be relevant (i.e. no incoherent topics and all the topics are interpretable) and eventually retained. Interpretation and labelling of the topics were done manually, based on the top 30 terms (Table 1).

Table 1: description of topics(*)

#Topic	Definition	Domain	Weight %
1	Development of immunity during early life in relation with breast feeding and the mode of delivery. Allergies in children.	Clinical	4.5
2	Cellulose digestion in ruminants and termites	Ecology & environment	2.2
3	Methods and techniques of analysis of fecal and gut microbiota	Techniques	4.9
4	IBD: bacterial composition, bacteriotherapy & faecal transplantation	Clinical	4.9
5	Food shelf life, food spoilage microbiota, food safety, fermented food technologies	Applied	3.9
6	Animal farming, animal health, feed & growth in relation with microbiota	Applied	3.5
7	Gut-brain axis, behavior, mental disorders, neurocognitive disorders	Clinical	6.7
8	Aquaculture, fish microbiota	Applied	2.1
9	Characterization of human microbiota: metagenomics and bioinformatics, taxonomic and functional diversity, ecological and evolutionary aspects	Basic	8.2
10	Antibiotic resistance, antimicrobial resistance, pathogenic microbiota	Clinical	4.7
11	Soil and plants microbial communities, biodegradation, bioremediation	Ecology & environment	5.7
12	Prebiotics, dietary fibers and dietary modulation of microbiota	Nutrition	4.5
13	Mutual interaction of phenolic compounds and gut microbiota	Nutrition	4.2
14	Airway microbiota	Clinical	2.6
15	Probiotics. Diet therapy and prevention	Nutrition	3.9
16	Pathogenesis of chronic liver diseases, gut-liver axis	Clinical	2.4
17	Obesity and its metabolic complications (metabolic syndrome, diabetes, lipid metabolism)	Clinical	7.9
18	Evolutionaty aspects of endosymbiosis. Experimental models of aging	Basic	3.1
19	Bacterial metabolism. Bacterial functional and comparative genetics, proteomics, glycomics	Techniques	2.4
20	Pathogenesis of IBD: Innate immunity, mucosal inflammation and gut microbiota	Clinical	8.0
21	Gut microbiota & cancerogenesis	Clinical	3.1
22	Oral microbiota	Clinical	4.4
23	Vaginal microbiota, HIV, STD, pregnancy, allergies in children	Clinical	2.3

(*) *LDA visualisation of topics available on:* https://figshare.com/s/8c42766c82de60b3b9ac

Overall, 11 of 23 topics refer to clinical medicine; their overall weight exceeds 50%. Most of them cover the pathological conditions associated with the gut microbiota (1, 4, 7, 16, 17, 20, 21) such as inflammatory bowel disease (IBD), cancer, metabolic syndrome, etc. Some clinical topics correspond to microbiotas of other bodily sites as vagina (23), respiratory tract (14) and oral cavity (22). Three topics deal with applied problems as animal farming, food safety and aquaculture (5, 6, 8). Three other topics are related to human and animals' nutrition (12, 13, 15); they encompass subjects of probiotics, prebiotics and phenolic compounds and account for about 13% of papers. Two topics (3 and 19) focus mostly on laboratory techniques. Two topics (9 and 18) addressing the evolutionary and experimental issues are connected to basic research and contribute to about 11% of papers. Topics 2 and 11 bear upon ecological and environmental aspects, which can be can be thought of as foundational for the microbiota concept.

It should be pointed out that some topics appear more stable than others: 14 topics stably emerged in the runs 22 to 25. By this is meant that their vocabulary is sufficiently specific to form distinct aggregations whatever the number of topics. These highly specific topics include oral microbiota, soil and plants microbial communities, antibiotic resistance, food safety & food spoilage microbiota, interaction of phenolic compounds and gut microbiota, obesity and its metabolic complications, pathogenesis of inflammatory bowel disease. On the contrary, some topics found to be sensitive to changes in the number of topics and tend to reorganise.

Figure 4 represents the evolution of popularity of topics over time (the timeline was truncated for better readability, since the publication number in 1999-2008 was comparatively too low). The rank of different topics and the widths of streams at each time point indicate the relative importance of topics in the research landscape.

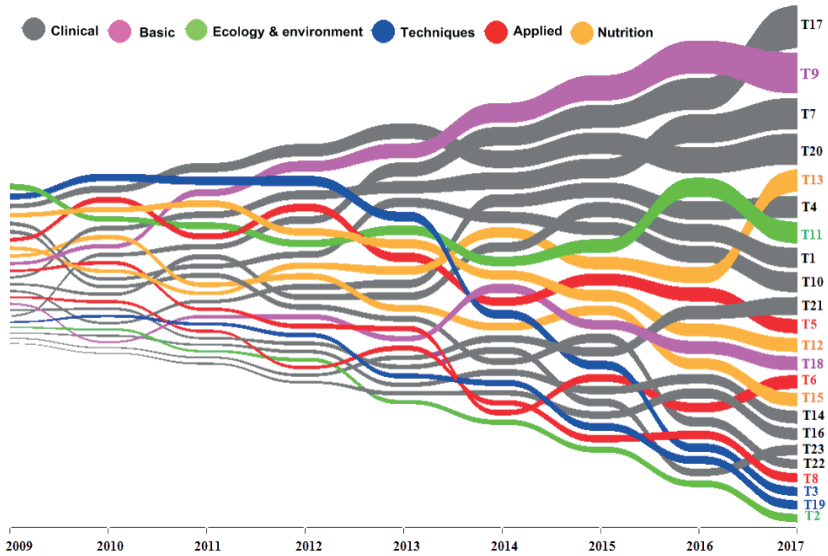


Figure 4: Evolution of topics weight by year

We can distinguish among the most precursory topics Soil and plants microbial communities, biodegradation, bioremediation (11) progressively losing its position despite the increase in output. Whereas in 2009 this topic represented more than 9%, at the end of the timeline it concerns only 5% of papers in the corpus. Interesting to note, the topic Methods and techniques of analysis of fecal and gut microbiota (3) which was one of the most popular topics in 2009-2012, displays a dramatic downturn and is found dropped to the bottom in 2016-2017.

Knowledge in microbiology allows more detailed interpretation of these observations. Thus, it may be hypothesised that this dynamics reflects the extensive surge of research activity driven by the revolution in screening technologies enabling once unfeasible compositional and functional analyses of complex bacterial communities. This period may be characterised as maturation of technologies and accumulation of empirical data. Beginning from 2012-2013 the clinically-oriented topics emerge and gain popularity. Obesity and its metabolic complications (17), Metagenomic characterisation of human microbiote (9), Gut-brain axis (7) and Pathogenesis of inflammatory bowel disease (1) become clearly dominant. It suggests that the microbiota research enters a translational phase and is consistent with the growing application of accumulated evidences to medical purposes. It can be inferred that the "community" or "biota" concept originated from ecology led to the paradigm shift in the way to consider the human-dwelling microbes as a part of ecosystem. This new viewpoint brings to deeper understanding of the complex relationships between the microbiota, health and disease. The succession of research focuses over time illustrates how diverse new findings contribute to advancement of knowledge on association of microbiota with pathogenesis of different health issues.

Figure 5 shows the scientific specialisation index by country. Colour degradation depends on the country's specialisation. The darker is the colour, the more the country is specialised in the topic. The degradation goes from yellow to red, through orange. In addition to the *heat map* representation, we have classified countries according to their specialisation index using Agglomerative Hierarchical Clustering (AHC) method. According to their specialisation, the countries are organised into three groups. The countries of the first group (Brazil, India, Spain, and Belgium) can be characterised by their specialisation in applied topics and under-specialisation in clinical topics. The countries of the second group (China, Korea, Japan, France, Germany, Italy, Canada, Australia and USA) tend to specialise in clinical topics. The third group comprise northern European countries, Ireland and Switzerland which show sporadically distributed specialisation, with lower-than-average activity in several topics.

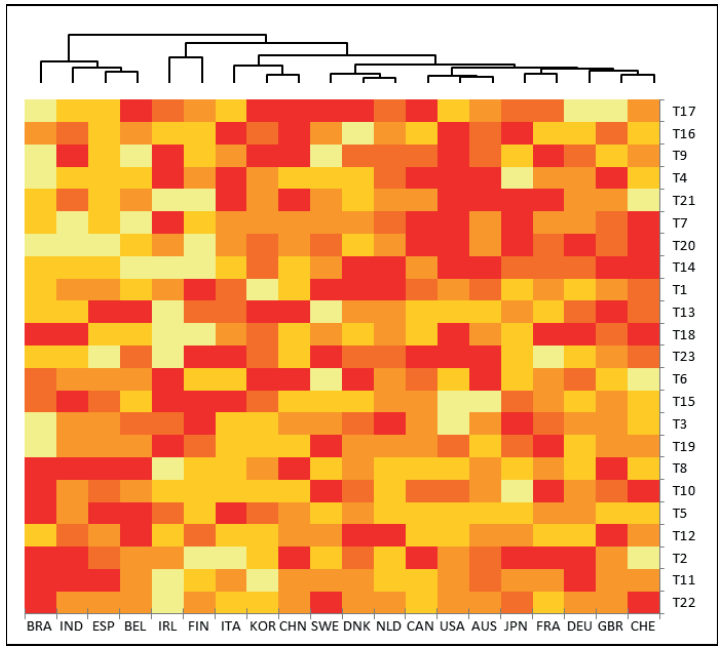


Figure 5: Heat map and Agglomerative Hierarchical Clustering by country, according to their specialisation index (by topic)

A closer examination reveals that the countries of the second group are members of the International Human Microbiome Consortium (IHMC, http://www.human-microbiome.org/), joint initiative aimed at coordination of research on human microbial communities, particularly through implementation of standard procedures and protocols and data sharing. This empirical finding suggests that the global-scale collaborative projects may be powerful tool shaping the countries' specialisation profiles.

Discussion and conclusion

The exponential growth of publications dealing with microbiota reflects a particular interest of the scientific community in the subject and reveals its great research potential. The growing body of literature contributed to an important progress in the knowledge of the microbiota's role in human health. Many nationally and internationally funded microbiota projects catalysed the publications output of the last decade (Hadrich, 2018; The Human Microbiome Project Consortium, 2012; *The Microbiome, diet and health*, 2017; The NIH HMP Working Group et al., 2009).

Since the turn of the millennium, the annual number of microbiota publications has been multiplied by 40. Distribution of countries' scientific activity in the field shows specific dynamics. The United States maintained their first place in terms of publications output between 2009 and 2017. The high reactivity of several countries is noteworthy: China experienced a major growth and moved up from 11th to 2nd position and Italy from 8th to

3rd. These examples witness to their capacities to rapidly leverage resources and to assert themselves on the world research forefront.

The microbiota research presents complex interdisciplinary field with topic structure changing over time. Originating from ecological sciences, it transitioned the period of technology maturation and data accumulation through large-scale screening and cataloguing. The recent focus on fast-growing clinically-oriented topics suggests that itentered the translational phase.

The current revolution in microbiota research is technology-driven as it is mainly due to high-throughput sequencing and to new culturing and stem cells technologies (Arnold, Roach, & Azcarate-Peril, 2016); it heavily relies on computing power to keep up with the growing volume of data. The advent of technology redefined research approaches in the field and strongly influenced the nature of scientific activities. Given the high costs of metagenomics and data processing, the microbiota research benefited from rapid internationalization. Therefore, the specialization of countries by topic is strongly related to funding for international research programs and international collaboration. Our findings show that the countries members of International Human Microbiome Consortium develop similar patterns of strong specialization in clinical research topics.

Our study analyses publications corpus built using solely the term “microbiota”. Covering the microbiota research in a more comprehensive way, with more exhaustive search query, could be a challenge for future studies. In addition to the topic structure of field and specialisation landscape it would be challenging to investigate the citation impact, to identify the major institutional players in the field, the collaboration patterns, etc.

This study illustrates the evolution of the concept underlying the complex interdisciplinary field and highlights the temporal changes of its topic structure and the specialisation profiles of different countries.

Bibliographic references

Arnold, J. W., Roach, J., & Azcarate-Peril, M. A. (2016). Emerging Technologies for Gut Microbiome Research. *Trends in microbiology*, *24*(11), 887–901. Retrieved January 25, 2019, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5074899/

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, *3*(4–5), 993–1022.

Blei, D. M. (2012). Probabilistic Topic Models. *Communications of the ACM*, *55*(4), 77–84.

Chang, J., Gerrish, S., Wang, C., Boyd-graber, J. L., & Blei, D. M. (2009). Reading Tea Leaves: How Humans Interpret Topic Models. *Advances in Neural Information Processing Systems*, 22, 9. Retrieved from https://papers.nips.cc/paper/3700-reading-tea-leaves-how-humans-interpret-topic-models

Garrido-Cardenas, J. A., Polo-López, M. I., & Oller-Alberola, I. (2017). Advanced microbial analysis for wastewater quality monitoring: metagenomics trend. *Applied Microbiology and Biotechnology*, *101*(20), 7445–7458. Retrieved September 28, 2018, from http://link.springer.com/10.1007/s00253-017-8490-3

Guo, H., Weingart, S., & Börner, K. (2011). Mixed-indicators model for identifying emerging research areas. *Scientometrics*, *89*(1), 421–435. Retrieved January 25, 2019, from http://link.springer.com/10.1007/s11192-011-0433-7

Hadrich, D. (2018). Microbiome Research Is Becoming the Key to Better Understanding Health and Nutrition. *Frontiers in Genetics*, *9*. Retrieved January 2, 2019, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6009232/

Lackey, J. B. (1956). Stream enrichment and microbiota. *Public Health Reports (Washington, D.C.: 1896)*, *71*(7), 708–718.

McCallum, A. K. (2002). *Machine Learning for Language Toolkit*. Retrieved from http://mallet.cs.umass.edu/

Observatoire des Sciences et Techniques. (2018). *La position scientifique de la France dans le monde 2000-2015* (p. 106). Hcéres. Retrieved January 2, 2019, from https://www.hceres.fr/content/download/32260/492585/file/rappScien_interieur_VF_BAT_web2b.pdf

Sievert, C., & Shirley, K. (2014). LDAvis: A method for visualizing and interpreting topics. *Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces* (pp. 63–70). Presented at the Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces, Baltimore, Maryland, USA: Association for Computational Linguistics. Retrieved December 20, 2018, from http://aclweb.org/anthology/W14-3110

The Human Microbiome Project Consortium. (2012). A framework for human microbiome research. *Nature*, *486*(7402), 215–221. Retrieved September 28, 2018, from http://www.nature.com/articles/nature11209

The Microbiome, diet and health: Towards a science and innovation agenda (OECD Science, Technology and Industry Policy Papers No. 42). (2017). Retrieved January 29, 2019, from https://www.oecd-ilibrary.org/science-and-technology/the-microbiome-diet-and-health_d496f56d-en

The NIH HMP Working Group, Peterson, J., Garges, S., Giovanni, M., McInnes, P., Wang, L., Schloss, J. A., et al. (2009). The NIH Human Microbiome Project. *Genome Research*, *19*(12), 2317–2323. Retrieved December 27, 2018, from http://genome.cshlp.org/cgi/doi/10.1101/gr.096651.109

Tian, J., Li, M., Lian, F., & Tong, X. (2017). The hundred most-cited publications in microbiota of diabetes research: A bibliometric analysis. *Medicine*, *96*(37), e7338. Retrieved September 26, 2018, from http://Insights.ovid.com/crossref?an=00005792-201709150-00002

Yao, H., Wan, J.-Y., Wang, C.-Z., Li, L., Wang, J., Li, Y., Huang, W.-H., et al. (2018). Bibliometric analysis of research on the role of intestinal microbiota in obesity. *PeerJ*, *6*, e5091. Retrieved September 26, 2018, from https://peerj.com/articles/5091

Zhang, J., Yu, Q., Zheng, F., Long, C., Lu, Z., & Duan, Z. (2016). Comparing keywords plus of WOS and author keywords: A case study of patient adherence research. *Journal of the Association for Information Science and Technology*, *67*(4), 967–972. Retrieved December 18, 2018, from https://onlinelibrary.wiley.com/doi/abs/10.1002/asi.23437

The presence and issues of altmetrics and citation data from Crossref for working papers with different identifiers from Econstor and RePEc in the discipline of Economic and Business Studies

Kaltrina Nuredini[1] and Isabella Peters [1](2)

[1]*k.nuredini @zbw.eu*
[1]ZBW – Leibniz Information Centre for Economics, Kiel, (Germany)

[1]*i.peters@zbw.eu*
[1]ZBW – Leibniz Information Centre for Economics, Kiel, (2) Kiel University, (Germany)

Abstract

For the past years, preprints started to be very common in Economics and Business Studies and economic researchers simply referred at them as working papers (Cruz & Krichel, 2000). Since a previous study of Nuredini & Peters (2016) confirms a relatively good coverage of journal articles for Economic and Business Studies literature in Altmetric.com, this study explores the altmetric representations for working papers in these fields. We present altmetric information from Altmetric.com for working papers from Econstor . Considering that working papers in Economics and Business Studies from Econstor often have handles for their identification, our study explored handles and confirmed a lower coverage in Altmetric.com (0.2%). Therefore we investigated altmetric information for two other working papers identifiers: DOIs, and URLs. A better coverage is identified for working papers with DOIs (7%). Econstor URLs are less found in Altmetric.com with coverage of 0.03%. The top most used altmetric source for working papers in Economic and Business Studies is Twitter for handles and DOIs and for URLs is Policy Posts. Mendeley counts are well present for working papers with DOIs but not for handles. A negative correlation (r = - 0.0157) is identified between citation counts from Crossref and Altmetric Scores. Additionally, we noticed several issues that are happening while sharing Econstor working papers on social media that prevent from collecting altmetric information. Thus, we suggest an alternative way to share these papers on social media to prevent losing altmetric information.

Introduction

Nowadays, with the use of digitization of the archives, there is an immense rise in disseminating research online (Speidel & Spitzer, 2018). Especially this rise effects open access repositories (e,g., arXiv[1], bioRxiv[2], OSF preprints[3] etc.) that host documents prior to formal publication or so called preprints (Speidel & Spitzer, 2018). According to Tomaiuolo & Packer (2000) preprints can be of different types: 1) papers that are not yet submitted to any journal, 2) papers that are under a peer-review process and waiting for publication decision and 3) papers that are electronically available that might fall in the category 1) and 2) or that can be used to assemble online feedback before submitting to a journal. Publishing a preprint provides various benefits to authors and readers such as 1) the research findings are published quickly and indexed in different services such as Google Scholar and Altmetric.com while traditional papers that are under review take longer (months or years) to be published and indexed; 2) Authors can collect feedback and further revisions about their research prior to a formal submission; and 3) papers that have a preprint gain more visibility and are 30% more cited than papers without a preprint (Speidel & Spitzel, 2018; Ozler, 2011; Krugman, 2013).
Preprints are very common in different fields of study i.e., physics (Delfanti, 2016) and life sciences (Serghiou, et al., 2018) as well as in Economics (Cruz & Krichel, 2000; Ozler, 2011). Preprints in Economics and Business Studies are simply referred to as working papers

[1] https://arxiv.org/
[2] https://www.biorxiv.org/
[3] https://osf.io/preprints/

(Frandsen, 2009; Ozler, 2011). Working papers are often issued from authors that are part of departments and research organizations rather than from individual authors (MacKie-Mason & Lougee, 2008). Working papers are published in informal series such as NBER[4], BREAD[5], World Bank Policy Research Working Paper Series[6] (Ozler, 2011) and as well as in repositories Econstor[7], SSRN[8], AgEcon Search[9] etc. One of the reasons that economists might have turned to working papers is because the journal review process for economic papers takes about two years followed by an extensive revision (Ellison, 2002). Publishing in economic journals seem to be the most time consuming tasks for economic researchers (Moffitt, 2011). So, according to Ellison (2002), working papers are used as means for disseminating information and journals are used to receive an approval of the quality of the paper.

Preprint repositories allow papers to be available online and that provides the opportunity to connect them with the audience through social media platforms (Shuai, Pepe & Bollen, 2012). All social media platform activities such as for example, tweets, shares on Facebook, mentions in blog posts, readership counts on Mendeley as well as mainstream media, downloads and policy posts are so called alternative metrics or altmetrics. Altmetrics may measure the societal impact of different research outputs (Wilsdon et al., 2015). Furthermore, altmetrics are known as a complimentary measure to traditional indicators (e.g., citations) for assessing impact (Costas et al., 2015). Similarly, as for other disciplines citations, seem to play an important role in economics as well, for reflecting the quality of the scholarly work of researchers and its achievements (Hammermesh, 2018).

Numerous studies showed that altmetric information significantly correlate with citations (Costas et al., 2015; Nuredini & Peters, 2016) suggesting that altmetrics may be related to scholarly activities anywise (Wilsdon et al., 2015). For example, a study found a significant correlation of social media mentions (i.e., tweets), downloads and citation counts for preprints in arXiv.org. The authors suggest that early tweets of preprints published in arXiv can lead later to higher download counts and more citation counts in arXiv (Shuai, Pepe & Bollen; 2012). A recent study of Bornmann & Haunschild (2018) explored altmetrics for papers in F1000Prime to determine whether altmetrics correlate with the scientific quality of papers. The authors found out that citations and readership counts from Mendeley are more related to quality as the tweets from Twitter. Another study of Nuredini & Peters (2016) found a relatively good coverage of journal articles for Economic and Business Studies literature in Altmetric.com. Serghiou at al., (2018) explored preprints in biological sciences for how they are used online via altmetric information and cited via the community. According to their results, they concluded that publications that had preprints have received more citations and altmetric scores as those publications without preprints.

Within this research, we will perform a study that looks at altmetric information and citation counts for working papers in Economic and Business Studies literature to investigate their visibility and impact within the online environment. According to Frandsen (2009) working papers seem not to be well cited especially in the field of economics and that the citation rate of working papers is similar to citation rates of low impact journals. This might happen because economic researchers are encouraged to wait for journal articles to read and cite rather than use working papers. Since working papers can be changed over time and can lead to different available versions (Ozler, 2011). Therefore, we would like to explore to what

[4] https://www.nber.org/papers/
[5] http://ibread.org/bread/papers
[6] http://www.worldbank.org/en/research/brief/policy-research-working-papers
[7] https://www.econstor.eu/about
[8] https://www.ssrn.com/en/
[9] https://ageconsearch.umn.edu/?ln=en

extent working papers are shared within social media platforms. Specifically, the main objective of this study is to identify the coverage of altmetric information for three different identifiers for working papers from Econstor and RePEc. Secondly, we will present their main issues with altmetric data and thirdly we will check whether working papers are preferably cited.
Accordingly, with this research study we would like to answer the following research questions:

1. What is the coverage of working papers of Economic and Business Studies literature for different identifiers form Econstor in Altmetric.com?
2. In which altmetric sources are working papers from Economic and Business Studies most often mentioned?
3. How is the citation rate from Crossref for working papers in Economic and Business Studies?

Methodology and data

This study selects working papers (with three different identifiers i.e., handles, DOIs, and URLs) for Economic and Business Studies from the Econstor repository and additionally considers handles and URLs from RePEc.
Econstor is a non-commercial and one of the largest repositories for scholarly economic literature. It includes different types of documents, such as working papers as well as journal articles and conference proceedings summing up to a total 165,000 resources that are freely accessible (Weiland, 2011). Around 100,000 documents in Econstor are working papers. Econstor as a disciplinary repository uses DSpace to assign identifiers respectively handles to working papers that officially started in 2009 (Borst & Weiland, 2009; Weiland, 2011). The handle[10] is a persistent identifier assigned to digital objects and other internet resources managed by the Handle System (Sun, Lannom & Boesch, 2003). Handle's first implementation system was developed in 1994 at CNRI[11]. Econstor is one of the main contributors to RePEc and one third of Econstor publications are also available in RePEc especially the content that comes from German institutions since Econstor is known as a "national input service" for RePEc (Weiland, 2011). RePEc is a decentralized database with over 2,000 archives from 99 countries that holds working papers and other research manuscripts in the discipline of Economics and Business Studies.
Altmetric.com[12] is queried for altmetric data and Crossref[13] for citations. Altmetric.com collects social media information for research products found online from specified sources such as social media platforms, traditional media, and online reference managers (Costas et al., 2015). It tracks eleven[14] different research identifiers such as DOIs, handles, RePEc IDs, URLs, ISBNs, SSRN IDs, PubMed IDs, arXiv, ADS IDs, URNs and ClinicalTrials.gov. Crossref[15] is a registration agency for scholarly communication; it provides metadata for DOIs and since 2017 offers free scholarly citation data for their own DOIs.
To perform our study for exploring altmetric information from Altmetric.com we selected Econstor working papers with three different identifiers: handles, DOIs, and URLs. Econstor

[10] http://www.handle.net
[11] https://www.cnri.reston.va.us/
[12] https://www.altmetric.com/
[13] https://www.crossref.org/
[14] https://help.altmetric.com/support/solutions/articles/6000134562-what-scholarly-identifiers-are-supported-by-altmetric-
[15] https://support.crossref.org/hc/en-us/articles/215787303-Crossref-Data-Software-Citation-Deposit-Guide-for-Publishers

also supports URNs[16] but we excluded them from this study since in Altmetric.com we couldn't find any records for URNs with altmetric information. Each entry in Econstor can have at least one of the identifiers listed above.
URLs in Econstor are represented in two different ways: a) *landing page* that present the page of metadata of a specified working paper and b) *pdf full text* that redirect to the .pdf of the document. Since our study focuses on working papers only, in Econstor 100,000 working paper are found for URLs with landing page and 100,000 URLs for pdf full text. Landing pages links start with a prefix (https://www.econstor.eu/handle/10419/) which is followed by an EconStor-local ID as the suffix. Pdf full-text links start with the prefix http://www.econstor.eu/bitstream/10419 and the Econstor ID follows as a suffix. Additionally, to Econstor URLs, we use also RePEc URLs (33,735) only for working papers that are listed in Econstor. RePEc URLs start with a prefix from the IDEAS website https://ideas.repec.org/p/zbw/ where "p" denotes the working paper and this URL is followed by the RePEc ID. The RePEc ID came un-encoded for example, one of the RePEc IDs is: RePEc:zbw:agawps:01 and to attach this ID at the URL we needed to encode into a URL form such as: zbw/agawps/01 and add .html at the end of the URL (https://ideas.repec.org/p/zbw/agawps/01.html).

Results

1: Coverage of working papers in Econstor and Altmetric.com

In Table 1 we present the coverage of working papers in Econstor, Altmetric.com, and Crossref additionally also working papers that are listed in Econstor and are available in RePEc. Dataset I contains working papers found in Econstor with handles - that sums up a total of 99,985 handles. We only selected the working papers with handles that have publication dates attached at them since we wanted to see the handle coverage of working papers in Econstor through the years. Dataset II has working papers with handles from Econstor that possess DOIs (4,605) and is a subset of Dataset I. Crossref is used only for Dataset II because Crossref API can be queried only by a DOI. Dataset III includes Econstor URLs of the working papers. Dataset IV includes all RePEc handles for working papers found in Econstor and Dataset V includes the RePEc URLs for working papers listed in Econstor.

Table 1: Datasets for working papers in Economic and Business Studies and their coverage in Econstor/RePEc, Altmetric.com and Crossref.

	Dataset I ***Handles only***		*Dataset II* ***Handles and DOIs***		*Dataset III* ***URLs***		*Dataset IV* ***Handles RePEc***		*Dataset V* ***URLs RePEc***
Found in Econstor	99,985		4,605		100,000		50,008		33,735
Found in Altmetric.com	246	0.25%	320	7%	37	0.03%	589	1.2%	0
Found in Crossref			1,613	35%					

[16] https://tools.ietf.org/html/rfc8458#page-13

Dataset I: Economic Working Papers with Handles Only

In Econstor, 100,000 handles that identify working papers for Economic and Business Studies are found (download date: 25.05.2018). Few papers do not have publication dates attached at them therefore, only 99,985 handles are retrieved with publication dates and upon them Dataset I is generated. Econstor records papers in two ways: 1) via internal full-service for paper series, journals or conferences, where the Econstor team organizes the full-text upload and metadata recording with no charge and 2) via self-archiving where single authors are able to self-archive their papers (i.e., pre/post publications, reports or thesis; Weiland, 2011).

According to the Figure 1, adding handles for working papers started to increase from 1998. Based on the publication year 2009 until 2016 we found roughly 50,000 handles assigned to working papers and that represents the half of coverage of the Dataset I.

On June 08th, 2018 the Altmetric Explorer[17] is used for downloading altmetric information and a MySQL database for holding and querying the dataset. The dataset I results in 661 handles found in Altmetric.com but only 246 handles have altmetric information with an altmetric score greater than 0[18]. The Altmetric attention score is a full number that indicates the attention the research output has received online. The score is calculated based on an algorithm provided from Altmetric.com that uses the weighted counts of each source used for (i.e., Twitter, news, blogs etc.) tracking research outputs[19].

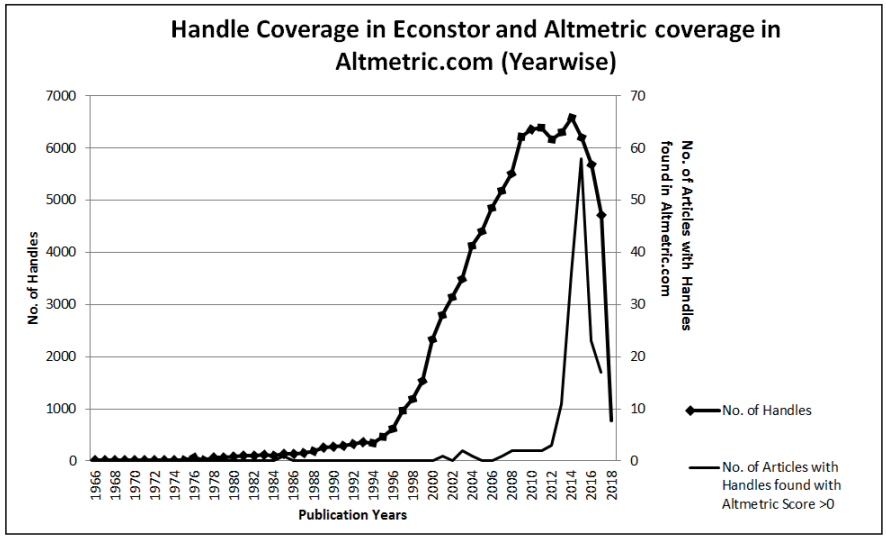


Figure 1: The coverage of Handles in Econstor and Altmetric.com.

The other rest of 415 handles have an altmetric score of 0 which means they are not tracked for any attention by Altmetric.com even though those handles are found in Altmetric.com. It is worth mentioning that Mendeley counts are not calculated in the altmetric score so there

[17] https://www.altmetric.com/explorer/outputs

[18] https://help.altmetric.com/support/solutions/articles/6000060970-putting-the-altmetric-attention-score-in-context

[19] https://help.altmetric.com/support/solutions/articles/6000060969-how-is-the-altmetric-attention-score-calculated-

might be working papers that have gained Mendeley attention but are not considered in the selection. Therefore from our Dataset I we found out only one handle that has Altmetric Score=0 but Mendeley count higher than 0 and we added this handle to our Dataset I.
The coverage of handles in Econstor and Altmetric.com is shown in Figure 1 where it only presents those articles that have Altmetric Score > 0 or Mendeley counts > 0 and had a publication date listed from Altmetric.com. Another limitation here appears because eighty-three articles out of 246 had no publication date from Altmetric.com and therefore they are not presented in Figure 1. Altmetric data for working papers are better present from the publication year 2011 and onwards since Altmetric.com started tracking attention in 2011. This leads to bias of altmetric information that also has been mentioned in other studies exploring other research products such as journal articles (e.g., Costas et al., 2015; Nuredini & Peters, 2016).

Dataset II: Economic Working Papers with Handles and DOIs

Given that the analysis of Dataset I with handles has a low coverage in Altmetric.com a second Dataset II was explored. For Dataset II we used DOI as an alternative persistent identifier for working papers in Economic and Business Studies. Thus, according to Dataset I we selected all the handles that have been assigned a DOI. We found a subset of Dataset I with 4,605 working papers with DOIs presenting Dataset II. From Figure 2 below, we can clearly see that DOIs are found mostly for handles that are published from 2009 until 2016. We assume that since 2009, DOIs appears to be more common in economic community for working papers and not only to journal articles.

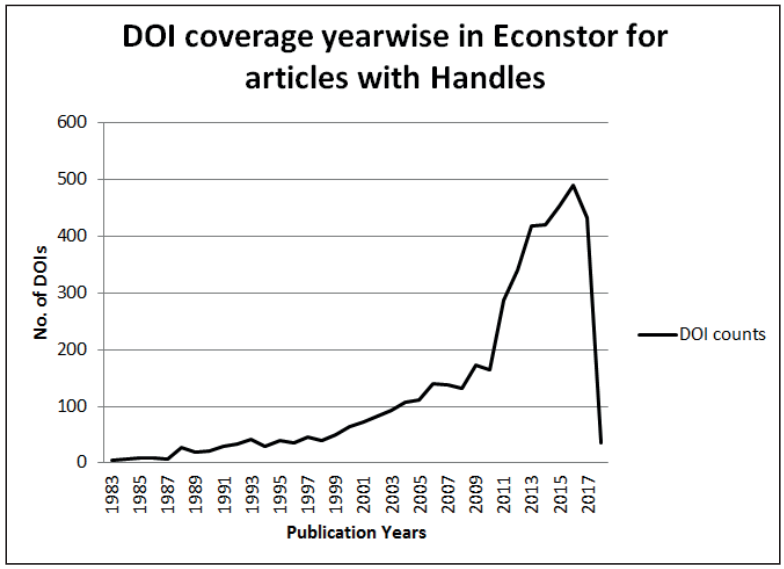


Figure 2: DOI coverage in Econstor for working papers with Handles.

Dataset II is found in Altmetric.com with 756 DOIs of which 244 DOIs have an Altmetric Score > 0 and 76 DOIs with an Altmetric Score = 0 but Mendeley counts > 0. This means that the total number of DOIs found with altmetric information is 320 and covers 7% of Dataset II.

Dataset III: Economic Working Papers with URLs

For dataset III Econstor URLs for identifying working papers in Economic and Business Studies are used. Econstor provides two types of links: *front-door* and *direct document links.* Both link types are searched in Altmetric.com for altmetric information. In Econstor 100,000 URLs are found for front-door links and 100,000 for direct document links.

a) *Landing page* links in Altmetric.com (date of study: 25.06.2018) are found only for 37 research outputs with altmetric information.
b) *Pdf full-text* links have no altmetric information in Altmetric.com (date of study: 25.06.2018).

According to a) the coverage of URLs is pretty low when it comes at sharing Econstor links especially via Twitter. For example, giving the Twitter news feed from NEP-DEV[20] they share almost every day at least one Econstor working paper link. NEP-DEV[21] is a twitter page tweeting the latest working papers from RePEc. Since Econstor is the main contributor to RePEc and one third of the content that is published in Econstor is also available in RePEc (Weiland, 2011) the coverage of URLs should have been higher in Altmetric.com. One reason that this low coverage of URLs appears in Altmetric.com is that NEP-DEV tweets short links which we assume that this feature makes it difficult from Altmetric.com to track the original links (i.e., landing page and pdf full-text links). The second reason might be that since Econstor contributes to RePEc there are cases (according to the tweet feeds) where for the Econstor working paper a RePEc front-door link is tweeted instead of an Econstor link. Therefore we created two new Datasets IV and V that looks up altmetric information for RePEc handles and URLs listed in Econstor.

Dataset IV and V: Economic Working Papers with Handles and URLs from RePEc listed in Econstor

Dataset IV includes RePEc handles (50,008) for working papers listed in Econstor. The handles are searched in Altmetric Explorer (date of study: 29.06.2018) and 589 are found, of which 583 have an altmetric score > 0. Thus, the coverage of RePEc handles is higher than Econstor handles in Altmetric.com.

Dataset V is composed of RePEc URLs and none of the URLs is found in Altmetric.com with altmetric information.

After exploring all the URLs in Altmetric.com from Dataset III and V we can conclude that short URLs make it difficult for Altmetric.com to track pdf full-text links shared in altmetric sources especially in Twitter. Specifically, Altmetric.com will not pick up a URL when it is shared for example in Twitter, but rather the directed page of that URL and then looks up for a handle, DOI or URL on that page. Moreover, the study of Zahedi & Costas (2018) that explores data collection from five (i.e., Altmetric.com, Mendeley, Crossref Event Data, Lagotto, Plum Analytics) altmetric aggregators claims that unfortunately there is no direct way to explore how altmetric aggregators query and collect altmetric information from third party APIs.

[20] https://twitter.com/search?f=tweets&q=econstor.eu%2Fbitstream%2F10419&src=typd
[21] https://twitter.com/RePEc_NEP_DEV

2: Best providers of altmetric sources

Dataset I

Table 2 below displays the top 6 altmetric sources for which altmetric information for the handles of Dataset I have been found. The calculation of total counts is done by selecting only articles that have altmetric score > 0 and summing up the usage numbers for every social media source tracked by Altmetric.com. Twitter is the source providing most altmetric counts for Economic and Business Studies working papers. Mendeley is not included in the Altmetric Score of Altmetric.com[22] , therefore, our selection criteria with an altmetric score > 0 excludes articles that only have gained Mendeley attention. News and Blogs have the same number of counts and because of that, we show 6 altmetric sources in Table 2. The total count of altmetric sources is 2,121 which is the total of each sum of the altmetric source received from all handles found in the dataset. The sum of the altmetric score for all handles found is 1,892. The difference between the sum of total counts from altmetric sources with the altmetric score is that the altmetric score[23] is calculated based on the weighted values each source has in Altmetric.com. And the sum of total counts from altmetric sources is provided by simple counts that these sources have received.

Table 2. Sum of Counts of Altmetric Sources for Handles in Altmetric.com.

Altmetric Source	**Total Count of Altmetric Sources**	**Coverage of handles in %**
Twitter	1,925	76%
Facebook	72	18%
Policy posts	50	15%
Google Plus	34	12
Blogs	20	6%
News	20	4%
Total	**2,121**	

Dataset II

Dataset II is the subset of Dataset I which includes only those records from Dataset I that have DOIs (4,605). Twitter is the most used altmetric source for working papers in Economics and Business Studies with DOIs. Mendeley, in this case, is the second source after Twitter even though for Dataset I the coverage of Mendeley was quite low. We assume that Mendeley users do not save working papers with handles but rather working paper with DOIs in their system.

The sum of the altmetric scores received for all papers searched in Altmetric.com via DOI is 1,946 which is lower than the sum of altmetric scores found via handles (2,121).

[22] https://help.altmetric.com/support/solutions/articles/6000060969-how-is-the-altmetric-attention-score-calculated-

[23] How altmetric score is calculated: https://help.altmetric.com/support/solutions/articles/6000060969-how-is-the-altmetric-attention-score-calculated-

Table 3. Sum of Counts of Altmetric Sources for Handles with DOIs in Altmetric.com

Altmetric Source	Total Count of Altmetric Sources	Coverage of DOIs in %
Twitter	1,136	58%
Mendeley	545	42%
Policy posts	170	27%
News	40	6%
Facebook	33	7%
Blogs	22	4%
Total	**1,946**	

Dataset III

The altmetric sources for URLs have a very low number of counts. Table 4 represents the top 3 altmetric sources for working papers with Econstor landing page URLs. Most of the working papers are mentioned in policy documents, news, and Wikipedia. Twitter, in this case, is not covered even though there is a high number of URLs shown. Policy documents in Altmetric.com are defined as reports, white papers or documents that provide policy and guidance from government or non-government organizations[24]. Thus, policy posts include references of working papers in policy documents. Altmetric.com searches for mentions in policy documents based on links, identifiers and text mining[25]. Text mining works by using a scraper that can match the mention in the policy document with an appropriate research output based on the author names, journal title and time frame. This step is needed when in the policy documents neither URL nor DOI is found. According to our results, we confirm that in the policy documents, working papers are mostly referenced by the URL. Same procedure is used in the News. Altmetric.com tracks around 2,900 news outlets via link recognition and news tracker mechanisms to pick up the mentions[26]. The news tracker mechanism is based on the search of the text of that news based on the author name and journals.

Table 4. Sum of Counts of Altmetric Sources for landing page URLs found in Altmetric.com

Altmetric Source	Total Count of Altmetric Sources
Policy posts	22
News	9
Wikipedia	4

Dataset IV

The altmetric sources for RePEc handles found in Econstor are explored. Table 5 presents top 3 altmetric sources of which Twitter has the highest altmetric score and Mendeley=0.

[24] https://help.altmetric.com/support/solutions/articles/6000060968-what-outputs-and-sources-does-altmetric-track-

[25] https://help.altmetric.com/support/solutions/articles/6000129069-how-does-altmetric-track-policy-documents-

[26] https://www.altmetric.com/about-our-data/our-sources/news/

Table 5. Sum of Counts of Altmetric Sources for RePEc handles listed in Econstor

Altmetric Source	Total Count of Altmetric Sources
Twitter	1,140
Facebook	91
Blogs	49

3: Working Paper citations from Crossref (Dataset II)

Dataset II with DOIs is used to query Crossref for citation data. 35% of DOIs are found in Crossref of which 231 DOIs (14%) have citation counts greater than 0 and 8% citation counts of 1. Citations from Crossref are based on the "is-referenced-by-count" parameter. According to our results, working papers with DOIs have relatively low citation counts. However, it is important to mention that Crossref has data limitations and can result in missing citations data.

Crossref calculates the citation counts upon the publishers that deposit reference lists. It is worth mentioning according to Crossref, that not all publishers deposit references lists and that not all references use DOIs. Therefore, there are limitations regarding the accumulation of Crossref citations that can result in low citation counts. Regardless of limitations in the citation data, we calculated Spearman rank correlation between altmetric score and citation counts for working papers that have DOIs. The Spearman rank correlation is calculated with $r = -0.0157$ and $p = 0.6746$; According to the results, there is no significant correlation between altmetric score and citation counts for working papers in Economics and Business Studies. This suggests that working papers that have high citation counts do not have high altmetric score and vice versa.

Discussions

With this study, we found out that altmetric information for working papers is differently presented for different identifiers. Altmetric information is better to present for working papers with DOIs (7%) rather than with handles or URLs. URLs either have relatively low coverage in Altmetric.com or no altmetric information at all. URLs are mostly presented to policy documents. We assume that the low coverage of URLs in Altmetric.com might fall in two categories: 1) might be a technical issue that is present when collecting altmetric information from different identifiers. Zahedi & Costas (2018) for example, mentioned that we should be aware of technical issues that appear while tracking different identifiers (shortened URLs, URLs, DOIs, PubMed, etc.) since they can influence the rate of altmetric information. For instance, URLs are often shared in Twitter according to the NEP-DEV Twitter stream. But most of URLs have no altmetric information. This might happen because Altmetric.com may not count URLs mentioned in a Twitter post but rather it checks for the URLs that might appear on the directed content page of the mentioned URL in Twitter. 2) Might be that working papers are not shared in social media because economists think that working papers are problematic to be shared. Ozler (2011) claimed that working papers in Economics can be changed - even significantly. Meaning that, the findings can be improved or changed over time before the original premise. Therefore he suggests that readers of working papers should be informed for every updated version of the working paper that might exist to avoid missing the new changes. However, he points out that since people are busy and they might not have the time to read every possible version of the working paper they are

interested in, more attractive for them to read are journal articles that contain the ultimate results.
Within this paper we also presented the technical issues that got in our way while accumulating altmetric information for economic working papers with different identifiers. To prevent technical issues mentioned above and to enable papers to gain more online attention via social media platforms (i.e., Twitter) we suggest to economists or other counterparts to tweet the handle or DOI of that particular paper additionally to the shortened link. With this approach, Altmetric.com can be able to record the altmetric information for shared working papers on Twitter without losing information.

Conclusions

This study explored to what extent Economic and Business Studies working papers are shared within social media platforms as well as showed the technical issues that were present during the process of data collection. Generally, for all working papers identifiers we selected, altmetric information is low. We assume that economic researchers do not often share working papers on their online social media platforms since working papers are often changing which leads to different working paper versions.
With this study we presented altmetric information for working papers in Economic and Business Studies from Econstor (additionally RePEc data listed in Econstor) using social media metrics from Altmetric.com. Our results are based on three different working paper identifiers and confirm different coverage of altmetric information. Our study confirms that working papers in Econstor are increasing especially recently. In Altmetric.com we found a better coverage of DOIs (7%) rather than for handles (0.2 %) and URLs from Econstor (0.03%).
The top most used altmetric sources are Twitter for working papers from Econstor for both identifiers handles and DOIs. For Mendeley, we found a lower coverage for working papers with handles, which means that not many researchers save handles in Mendeley but rather DOIs. RePEc handles, on the other hand, have a higher coverage with 1.2% as in comparison with Econstor handles.
According to the citation rate of working papers gathered from Crossref, we found out that working papers in Economics and Business Studies from Econstor are not well cited which confirms the study of Frandsen (2009). We indicate that working papers in Economic and Business Studies are not well shared online as well which confirms the study of (Ozler, 2011). We also show that Altmetric Scores and citation counts for working papers are not significantly correlated. Since citations received from Crossref are represented for a smaller number of articles and with a low citation rates it is worth mentioning that the citation dataset in this study has limitations. Therefore in order to validate the results more research is needed. Moreover, a complete citation database should be used to confirm the results of this study. Another limitation of this study is the coverage of working papers. The working papers investigated here are only based on Econstor titles and their identification numbers but do not include working papers published in other databases or series.

Acknowledgments

We would like to thank Jan Weiland from ZBW (Department: Publication Services) for his support in providing access to and explaining Econstor metadata. We also thank Altmetric.com for offering us the opportunity to access and use altmetric data.

References

Bornmann, L., & Haunschild, R. (2018). Do altmetrics correlate with the quality of papers? A large-scale empirical study based on F1000Prime data. *PloS one*, *13*(5), e0197133.

Borst, T., & Weiland, J. (2009). EconStor: ein fachliches Repositorium für die Wirtschaftswissenschaften. *GMS Medizin, Bibliothek, Information*, *9*(1), 2009-9.

Costas, R., Zahedi, Z., & Wouters, P. (2015). Do "altmetrics" correlate with citations? Extensive comparison of altmetric indicators with citations from a multidisciplinary perspective. *Journal of the Association for Information Science and Technology*, *66*(10), 2003–2019. doi: 10.1002/asi.23309 .

Cruz, J. M. B., & Krichel, T. (2000). Cataloging economics preprints: An introduction to the RePEc project. *Journal of Internet Cataloging*, *3*(2-3), 227-241.

Delfanti, A. (2016). Beams of particles and papers: How digital preprint archives shape authorship and credit. *Social studies of science*, *46*(4), 629-645.

Hamermesh, D. S. (2018). Citations in economics: measurement, uses, and impacts. *Journal of Economic Literature*, *56*(1), 115-56.

Ellison, G. (2002). The slowdown of the economics publishing process. *Journal of political Economy*, *110*(5), 947-993.

Frandsen, T. F. (2009). The effects of open access on un-published documents: A case study of economics working papers. *Journal of Informetrics*, *3*(2), 124-133.

Krugman, P. (2013, April 22). Understanding the NBER. [Blogpost]. Retrieved from https://krugman.blogs.nytimes.com/2013/04/22/understanding-the-nber/

MacKie-Mason, J. K., & Lougee, W. P. (2008). *Economics and Usage of Digital Libraries: Byting the Bullet*. Michigan Publishing, University of Michigan Library.

Moffitt, R. (2011). Getting published in Economics journals. Newsletter of the committee on the status of woman in the economics profession, (4-5).

Nuredini, K., & Peters, I. (2016). Enriching the knowledge of altmetrics studies by exploring social media metrics for Economic and Business Studies journals. In *Proceedings of the 21st International Conference on Science and Technology Indicators (STI Conference 2016), València (Spain), September 14-16, 2016*. Berlin: European Network of Indicator Designers (ENID).

Ozler, B. (2011, June 30). Working Papers are NOT working. [Blog post]. Retrieved from http://blogs.worldbank.org/impactevaluations/working-papers-are-not-working

Serghiou, S., & Ioannidis, J. P. (2018). Altmetric Scores, Citations, and Publication of Studies Posted as Preprints. *JAMA*, *319*(4), 402-404.

Speidel, R., & Spitzer, M. (2018). The What, the Why, the How. [Blog post]. Retrieved from https://cos.io/blog/preprints-what-why-how/

Sun, S., Lannom, L., & Boesch, B. (2003). *Handle system overview* (No. RFC 3650).

Shuai, X., Pepe, A., & Bollen, J. (2012). How the scientific community reacts to newly submitted preprints: Article downloads, twitter mentions, and citations. *PloS one*, *7*(11), e47523.

Tomaiuolo, N. G., & Packer, J. G. (2000). Pre-print servers: pushing the envelope of electronic scholarly publishing. *Searcher*, *8*(9), 53.

Weiland, Jan (2011). Econstor: A RePEc archive for Research in Germany. [Blog post]. Retrieved from https://blog.repec.org/2011/09/15/econstor-a-repec-archive-for-research-from-germany/

Wilsdon J, Allen L, Belfiore E, Campbell P, Curry S, Hill S et al. (2015) The Metric Tide: Report of the Independent Review of the Role of Metrics in Research Assessment and Management. Bristol, UK: Higher Education Funding Council for England (HEFCE).

Zahedi, Z., & Costas, R. (2018). General discussion of data quality challenges in social media metrics: Extensive comparison of four major altmetric data aggregators. *PloS one*, *13*(5), e0197326.

Conceptualising dimensions of bibliometric assessment: From resource allocation systems to evaluative landscapes

Fredrik Åström[1] and Björn Hammarfelt[2]

[1] *fredrik.astrom@ub.lu.se*
Lund University, Lund University Library, P.O. Box 3, 22100 Lund (Sweden)

[2] *bjorn.hammarfelt@hb.se*
University of Borås, Swedish School of Library and Information Science, 501 90, Borås (Sweden)

Abstract

The purpose of this paper is to discuss the conceptualisation of bibliometric analyses in terms of the levels on which they are performed, adding contextual factors to the dimension where the size of the unit being analysed is considered. Based on empirical investigations of resource allocation systems and research evaluation practices, as well as the previous literature conceptualising bibliometric analyses, a framework based on Whitley's (2000) notion of research fields as 'reputational work organisations', is discussed. The results suggest adding a contextual 'reputational dimension' to the size-based dimension distinguishing between micro-, meso- and macro-level analyses. Furthermore, we propose that 'evaluative landscapes' (Brandter, 2017) might be a fruitful approach for further analysing how complex and multifaceted landscapes of research assessment affects the individual researcher.

Introduction

In bibliometrics and other indicator based research evaluation and funding allocation practices, a distinction is typically made between micro-, meso- and macro-level analyses (Vinkler, 1988). Albeit functional, it is however, a classification that can be perceived as one-dimensional, taking into account only the size of the unit/s being analysed; not the context in which the analysis is being performed. It can be argued that contextual factors such as the purpose of the evaluation should be considered when classifying bibliometric or indicator driven analyses, as well as taking into account if the origin of evaluation criteria for a specific situation are determined by local or national institutional policies, or if they rather relate to criteria developed within specific fields of research. Conceptualising dimensions of research evaluation, bibliometric analyses and PRFS may seem as something of an esoteric exercise, but apart from ambitions towards conceptual clarity and nuance; the conceptual issues are also a part of a wider discussion on how to navigate a situation where there is a plethora of evaluation criteria coming out of various contexts related either to institutional policies or field related assessments of reputation, at the same time as the conditions for individuals or institutions to relate to – or negotiate – the criteria varies greatly, depending on for instance the level of being established within a specific system of research (Brandtner, 2017).

The aim of this paper is to discuss the classification of bibliometric analyses in terms of levels on which the analyses are performed; to attempt a conceptualisation of analytical levels taking into account contextual dimensions other than the size of the unit analysed. The conceptual discussion is in turn related to questions on: the relation between institutional and field related evaluation criteria; the possibilities of navigating within a complex system of evaluative situations; and, how this can be understood as an 'evaluative landscape' as formulated by Brandtner (2017).

Methodological considerations

This paper is a conceptual discussion of the categorisation of bibliometric analyses in terms of the classification of levels of analysis. Thus, the specific unit of analysis for this paper is the

concepts per se, and how they have been discussed in the literature. The concepts have been analysed through a close reading of literature concerning the conceptualisation of analytical levels in bibliometrics, which in turn has been contextualised through literature concerning research evaluation and PRFS practices.

The analysis of the literature on analytical levels in bibliometrics and evaluation and funding allocation practices is informed by the theoretical framework proposed by Whitley (2000), who describes research fields as 'reputational work organisations', with a particular focus on Whitley's concepts 'mutual dependency', the extent of which researchers are dependent on colleagues and other groups to make significant contributions to the field and thereby gaining reputation and merit; and 'reputational autonomy', the degree of a research fields control over standards for assessing competence and performance. In addition we make use of Brandtner's (2017) conceptualisation of 'evaluative landscapes' and discuss what such an approach might offer in terms of understanding bibliometric evaluation as part of a broader context of assessment.

Conceptualising levels of analysis in bibliometrics, research evaluation and PRFS

Typically, in terms of levels of application, bibliometric analyses are categorised through the size of the unit of analysis. Analyses of individual persons or research groups are considered as micro-level analysis. If the unit of analysis is for instance a department at a university, the analysis is considered to be on the meso-level. And if the unit is a higher education institution (HEI) such as a university, or even larger units such as countries or other geopolitical regions, the analysis is considered to be on the macro-level (Vinkler, 1988). This categorisation covers all kinds of bibliometric analyses, whether performed for evaluating the outcomes of a research project or the output of a university; or if performed for the purpose of allocating funds between departments at a local faculty or between HEIs in a country based on previous performance. This categorisation makes it possible to distinguish the use of bibliometrics in for instance, on one hand, the assessment of a research group's performance when evaluating a funding proposal as a micro-level analysis; from on the other bibliometrics being used to allocate research funds between HEIs in a country as a macro-level analysis. However, it does not make it possible to distinguish the analyses as on one hand, being part of an evaluation based on criteria coming out of a specific research field, as in the case of bibliometrics being part of a peer review assessment of a funding proposal; or on the other hand, an assessment of research performance based on indicators decided upon in an institutional policy process.

Empirical investigations of bibliometrics based research evaluation practices and PRFS

In terms of PRFS, the use of bibliometrics based allocation systems is well documented, both in terms of national systems (e.g. Hicks, 2012; Jonkers & Zacharewicz, 2016), as well as the extent to which PRFS have been applied locally (Aagaard, 2015; Hammarfelt et.al, 2016). In terms of analytical levels, as defined by for instance Vinkler (1988), we find examples of PRFS operating on all levels, from local PRFS distributing resources in-between individual scholars (micro-level) and departments or faculties (meso-level), to national systems distributing funds in-between HEIs based on bibliometric analyses (macro-level). These systems all come out of institutional policy decisions; and more often than not, does not take particularities of different research fields into account, more than possibly trying to identify means of balancing field differences when developing the indicators; and thus, reflects the goals or intentions of a local institution or national research systems, rather than for instance

the goals of a research field, that to various degrees are defined either internationally or on a more local level.

However, research evaluation and decisions on access to research funds and resources are also done on individual basis by peers of the scholars/scientists or research groups; and here, indicators of different kinds are used in support of the assessment – a practice sometimes referred to as 'informed peer review' (Abramo & D'Angelo, 2011) or 'citizen bibliometrics' (Hammarfelt & Rushforth, 2017; Leydesdorff et.al, 2016). Such practices has for instance been studied through analyses of peer assessment reports ranking candidates for academic positions (Hammarfelt, 2017; Hammarfelt & Rushforth, 2017), instructions for – and assessments of – applications for the Swedish academic title of 'docent'/Reader (Joelsson et.al, forthcoming), and through questionnaires (Hammarfelt & Haddow, 2018). What is apparent in these studies, is how the definition of assessment criteria in these situations are related to the research fields of the applicants rather than institutional contexts, even though both the assessment criteria formulated for the decision or the promotion, and the decisions made on basis of the assessments, are local institutional matter. The assessors are typically external peers of the applicant in the same field, thus incorporating assessment criteria valid for the respective research field, rather than the institutional environment. Thus, the 'reputational', and disciplinary, influence over resources is considerable and may come in conflict with organisational control. The assessment criteria varies between fields, how publications are assessed, and not the least in terms of the use of indicators; to what extent they are being used at all, but also what kind of indicators are important for assessing academic merit. At the same time, it also important to note differences between research fields – and how the contribution to the field by individuals is assessed – in terms of to what extent they can be considered international or local. While some fields have assessment criteria that are essentially the same independent of research environment, some fields show variation in assessment criteria depending on national research systems (Hammarfelt & Haddow, 2018).

Thus, we have two evaluative spheres. The institutional, where assessment criteria are formulated locally or nationally, related to policies and/or organisational goals; but where there also may be great variations in terms of indicators or evaluation criteria, between national and local PRFS, in-between universities, as well as between for instance faculties at individual universities. The other sphere of evaluation is the one based on peer assessment, where the criteria for assessment are primarily coming out of how academic merit is assessed within different research fields; albeit with varying degrees of local variations depending on the field. Both within and in-between these two overlapping spheres, there may be substantial variations in terms of criteria being used, some of which might even be directly conflicting. The question is how this plurality of both evaluating bodies and evaluation practices affect individuals and organisations; and how to balance various evaluation criteria? Brandtner (2017) describes the plurality as 'evaluative landscapes', and goes on to describe how this can have a compromising effect on the homogenising effect of evaluative practices on organisational behaviour.

Re-conceptualising levels of analysis in bibliometrics, research evaluation and PRFS

To conceptualise the evaluative and resource allocation contexts, and re-conceptualise the notion of levels of analysis in bibliometrics, the attention is turned to Whitley (2000), describing research fields as 'reputational work organisations', where the assessment of academic merit is an essential part in both the definition and organisation of a field. One important aspect of this is the concept of 'mutual dependency', that is how dependent scholars

and scientists are on their colleagues within the field for making significant contributions, including for instance the level of specificity of standards in terms of research practices and assessment criteria. Mutual dependency is described as one dimension through which research fields can be understood; and is related to another key dimension, that of 'task uncertainty', that is the degree of which research outcomes can be predicted based on aspects such as the formalisation of problem formulation, levels of standards in terms of methodological familiarity and so on. Another important aspect in Whitley's framework in relation to research evaluation practices is the concept of 'reputational autonomy, that is, to what extent a field autonomously control standards for assessing competence and performance; and following that, also the extent to which boundaries can be – or for that matter, are important to – maintain.

In Hammarfelt's (2017) comparison between biomedicine, economic and history research, we can see how in history and economics, reputational autonomy is maintained by a strong definition of boundaries by assessing applicants based on the field where they got their PhD degree or if the applicants have published articles in journals within the field. In biomedicine on the other hand, the background of the applicant – and the boundaries of the field in that sense – are less important. Instead, the autonomy of the field is rather maintained through a low level of task uncertainty combined with a high level of mutual dependency, regardless of the research field background of the researcher, the applicant is assessed by looking at merits in terms of the ability to perform very specific tasks using a set of well-defined methods. However, whereas the reputational autonomy of economics and history research – and in this case, they are also regarded as representatives of the social sciences and humanities (SSH) more generally – is maintained through the control over field boundaries by looking at from which field the applicants are coming from, some of the actual criteria for assessment – not the least in terms of to what extent publication related indicators play a role – seem more sensitive to developments outside the own field. While a field like biomedicine have a strongly defined set of criteria to assess those contributing to the field, and criteria substantially defined within the field; in SSH, assessment criteria seem to have a stronger relation to outside factors such as PRFS on different levels, as reflected in changes in publication patterns in SSH as exemplified by for instance Sīle & Vanderstraeten (2019) and Guns et.al (2019).

Field differences as interpreted through the framework of Whitley may primarily be seen as associated with evaluations and assessments done within the field, in economics and history primarily reflecting the pursuit of reputational autonomy, in biomedicine more related to a high level of mutual dependency; and to a lesser extent in relation to assessments of performance related to PRFS. However, in local PRFS at levels below the whole university, such as at faculty or department level, there are variations in what kinds of indicators being used, taking into account publication patterns and assessment criteria in different fields (Hammarfelt et.al, 2016). And even in terms of national PRFS, differences between fields are being taken into account: either by developing performance indicators that try to balance the differences by giving different weights to different publications, such as in Norway; or by trying to change publication patterns – and thus assessment criteria – by rewarding certain types of publications, as was one the original motivations for the Swedish PRFS (Nelhans, 2013). Generally, though, PRFS – and other institutional policy related evaluations – are primarily related to Whitley's concept of reputational autonomy; by creating a system of reward and access to resources that is not coming out of assessment criteria developed within the research fields, but rather is formulated in institutional policy processes.

Thus we find – in Brandtner's (2017) terms – an evaluation landscape with assessments being done on all levels of aggregation, from individuals to HEIs, reflecting the dimension of size of unit of assessment as categorised by for instance Vinkler (1988). But we also find the dimension of evaluation practices and assessment criteria, either coming out of institution policy processes, or those related to specific research field practices in terms of how to decide on how someone has contributed/can be expected to contribute to the development of the field. Brandtner describes evaluation landscapes as "the collectivity of evaluation practices including rankings, ratings and rewards" (Brandtner, 2017, 201), not only describing what is going on within a field, but also being "external arbiters that motivate organizations to converge toward similar behavior." (Brandtner, 2017, 204) However, considering the wide range of variations of evaluation practices and use of different bibliometrics based indicators both in-between and within the two dimensions described here, we find what Brandtner (2017) would describe the academic evaluation landscape as characterised by plurality; where assessment criteria are less likely to inform organisational behaviour in a field. This may also be one of the reasons for why it is difficult to study the concrete effects of specific bibliometric models of evaluation (cf. Waltman 2017)

In terms of academic research in general, according to Brandtner, the influence of evaluation practices should be relatively minor. At the same time, the 'Whitleyan' concepts still need to be taken into account: the level of reputational autonomy and mutual dependency of a field have a strong influence on to what extent researchers within a field are dependent on being assessed by other actors within the field (with access to resources largely coming out of funding structures more or less related to the field), or if they rather are dependent on access to resources provided on an institutional basis (e.g. time for research within a faculty position). What, in Brandtner's terms, should be an environment less influenced by evaluation practices, becomes a balancing act between various – and varying – assessment criteria.

In conclusion

This paper suggests that, in the categorisation of bibliometric analyses, research evaluation systems and PRFS, to not only direct attention to the size of the unit being analysed, but to also add a dimension taking other contextual aspects into account: the 'reputational' dimension: where assessment criteria are related to the organisation of different research fields. In fact, it might be argued that PRFS and other evaluation systems and assessment procedures are better understood as part of a larger 'evaluative landscape' (Brandtner, 2017), in which individual researchers has to navigate in a complex setting of indicators and evaluation practices. A specific tension, describes in this paper, is the one between institutional evaluation practices, and the reputation system, which is mainly governed through discipline-based collegial norms. A further complicating factor is that an institutions or individual's specific role and position in this 'landscape' may render rather different interpretations of, and reactions to, assessments procedures and evaluation systems. For example, a renowned institution may respond rather differently to measurements or ranking lists compared to less established universities. Similarly, a distinguished professor might react to the demands of assessment systems and indicators rather differently compared to a young untenured researcher. Consequently, we see the navigation of a plural evaluation landscape, where there is a balancing act between evaluations on individual, organisational and national levels, as well as between institutional and reputational assessments; and where the extent to which assessment criteria coming out if research field practices or institutional policy practices influence organisational – and individual – behaviour depends on the reputational autonomy of the field, as well as the field's level of (within-the-field) mutual dependency.

Acknowledgments

This work was supported by Riksbankens Jubileumsfond: The Swedish Foundation for the Social Sciences and Humanities (SGO14-1153:1).

References

Aagaard, K. (2015). How incentives trickle down: Local use of a national bibliometric indicator system. *Science and Public Policy*, 42(5), 725-737.

Abramo, G., & D'Angelo, C. A. (2011). Evaluating research: from informed peer review to bibliometrics. *Scientometrics*, 87(3), 499-514.

Brandtner, C. (2017). Putting the world in orders: Plurality in organizational evaluation. *Sociological Theory*, *35*(3), 200–227.

Guns, R., Eykens, J. & Engels, T. (2019). To what extent do successive cohorts adopt different publications patterns? Peer review, language use and publication types in the social sciences and humanities. *Frontiers in Research Metrics and Analytics*, 3:38.

Hammarfelt, B. (2017). Recognition and reward in the academy: Valuing publication oeuvres in biomedicine, economics and history. *Aslib Journal of Information Management*, 69(5), 607-623.

Hammarfelt, B. & Haddow, G. (2018). Conflicting measures and values: How humanities scholars in Australia and Sweden use and react to bibliometric indicators *Journal of the Association for Information Science and Technology*, 69(7), 924-935.

Hammarfelt, B., Nelhans, G., Eklund, P. & Åström, F. (2016). The heterogeneous landscape of bibliometric indicators: Evaluating models for allocating resources at Swedish universities. *Research Evaluation*, 25(3), 292-305.

Hammarfelt, B. & Rushforth, AD. (2017). Indicators as judgment devices: An empirical study of citizen bibliometrics in research evaluation. *Research Evaluation*, 26(3), 169-180.

Hicks D. (2012). Performance-based University Research Funding Systems. *Research Policy*, 41(2), 251-261.

Joelsson, E., Nelhans, G & Helgesson, C-F. (forthcoming). *Hur värderas publiceringsmeriter i det svenska akademiska systemet? En undersökning av värderingen av befordran till docent med särskilt fokus på betydelsen av öppen tillgång [What is the value of publications in the Swedish academic system? A study of assessments for the promotion to 'docent' with a particular focus on open access]*. Stockholm: Kungliga Biblioteket.

Jonkers, K. & Zacharewicz, T. (2016). *Research Performance Based Funding Systems: a Comparative Assessment*. European Union: JRC Science for Policy Report (EUR 27837 EN).

Leydesdorff, L., Wouters, P. & Bornmann, L. (2016). Professional and citizen bibliometrics: Complementarities and ambivalences in the development and use of indicators — a state-of-the-art report. *Scientometrics*, 109(3), 2129-2150.

Nelhans, G. (2013). *Citeringens praktiker: Det vetenskapliga publicerandet som teori, metod och forskningspolitik*. Gothenburg: University of Gothenburg (Diss.).

Sīle, L. & Vanderstraeten, R. (2019). Measuring changes in publication patterns in a context of performance-based research funding systems: the case of educational research in the University of Gothenburg (2005–2014). *Scientometrics*, 118(1), 71-91.

Waltman, L. (2017). Special section on performance-based research funding systems. *Journal of Informetric*s, 11(3), 904.

Vinkler, P. (1988). An attempt of surveying and classifying bibliometric indicators for scientometric purposes. *Scientometrics*, 13(5-6), 239-259.

Whitley R. (2000). *The Intellectual and Social Organization of the Sciences*. Oxford: University Press.

HEIs participations and mobility in the European Framework Programmes

Barbara Antonioli Mantegazzini[1] Benedetto Lepori[1]

[1] barbara.antonioli@usi.ch benedetto.lepori@usi.ch
Università della Svizzera Italiana, Via Lambertenghi 10A, Lugano (Switzerland)

Abstract

In this paper, we analyze changes in the participation of European HEIs to European Framework Programmes between 2008 and 2015. The goal is twofold: first, with the help of a transition probability matrix, we describe the mobility of HEIs in the considered period from and to a "core participant" network; a specific index will score that mobility. We then aim at identifying variables that help in explaining mobility of individual HEIs. In detail, we suggest that the probability of a change of the HEI's participation is influenced by endogenous variables such as individual characteristics, or by external ones, mainly country economics variable, different national research policies/strategies, as well as changes in the FPs' composition. Results represent the base for several public policies considerations and suggestions on how to strengthen the European Research Area.

Introduction

There is a vast literature on patterns of participation of Higher Education Institutions (HEIs) to European Framework Programs (EU-FPs), as well as on the factors influencing that level (see Lepori, Heller-Schuh, Scherngell and Barber 2014; Enger and Castellacci 2016). A common finding of this literature is the presence of a small, central core of HEIs with high levels of participation to the EU-FPs, which is lasting over time. This structure has been explained by the importance of relational ties in establishing European consortia, but also by cumulative effects (Merton 1988) related to the concentration of academic reputation at a few places in Europe. According to the EU-FPs policy goals, such a structure might be welcome because it concentrates resources and achieves excellence, but raises concerns in terms of cohesion of the European Research Area, given that the core is strongly concentrated in few countries, typically located in Western Europe. While the literature on the determinants of the participation is huge, the investigation on the institutional mobility with respect to their level of participation to EUFPs, respectively of entry and exit to the core of regular participants, is still limited. The analysis of factors and variables (endogenous and exogenous) that affect the probability of an HEI to change its membership to this participants' core could help the current debate, also supporting policy makers in their decision process. The available literature suggests a number of potentially relevant factors, including changes in the HEIs' resources (possibly as an outcome of changes at the national level in research strategies or financial policies), changes in the HEI's reputation (as the key factor associated with EU-FP centrality) and, finally, changes in the orientation of the programs, like the introduction with the Seven Framework Programme (2007-2013) of the excellence-oriented European Council Grants, ERC (Nedeva 2013, Luukkonen 2013).

Given the above considerations, the present contribution has a double objective: first, we want to describe and analyze this HEIs' mobility from and to the "core participants"; second, we aim at identifying variables that help in explaining changes in the number of participation. With reference to the latter, we speculate that the probability of a change of the HEI's EU-FP participation (measured with an index of class mobility) is influenced by endogenous variables such as individual characteristics, or by external ones mainly country economics variable, different national research policies/strategies and /or differences in the FPs' composition.

To this aim, the paper builds on a unique dataset providing data on EU-FP participations and on institutional characteristics for the years 2008 and 2015 developed as an extension of the European Tertiary Education Register (ETER; www.eter-project.com) developed within the EU infrastructure project RISIS (risis.eu).

Main literature and current debate

Research funding programmes such as FPs and Horizon2020 have registered an impressive rise in importance during the last decades, and this relevance has been reflected also in the academic debate. A large body of the literature has been devoted to the study of the main characteristics of the participation process and of the factors that may influence university participation in EUfunded R&D cooperative projects. The first line of analysis is dedicated to the investigation of the participation process. Main results point out that several characteristics influence HEIs' level of participation; scientific capabilities of applying institutions (publications and academic reputation) seem to be crucial (Geuna 1998; Seeber, Lepori, Montauti, et al 2015; Nokkala, Heller-Schuh and Paier 2011; Henriques, Schoen and Pontikakis 2009), as well as prior experience and relational capital are predominant factors in determining the involvement of HEIs (Nokkala, Heller-Schuh and Paier 2011). Project administration, administrative/bureaucratic effort seems to be relevant as well, especially for HEIs located in Eastern Europe or of small size (Mataković and Novak 2013), as well as lack of practice in competitive fund raising (Geuna 1998). Numerous high reputational organizations (mainly in UK) display a lasting high level of participation and assume a central position within the participation networks (Heller-Schuh, Barber, Henriques, et al 2011; Barber, Krueger, Krueger and Roediger-Schluga 2006;). Nokkala, Heller-Schuh and Paier 2011 investigated the influence of HEIs ranking position, which seems to be relevant only for project coordination. Large attention has been dedicated also to the analysis of the network structure of participations, adopting methods from social network analysis. Main results highlight an integrated and tightly-knit network (Scherngell and Barber 2011). Less experienced groups face barriers in entering in the EU competition; their participation remains marginal (Primeri and Reale 2012). A second line of analysis is aimed to the study of the exogenous conditions that influence participation. Results shows that the structure and the design of national funding strategies are relevant and interact with changes in EU funding policies (Luukkonen and Nedeva 2010). Also national economics matter: HEIs from richer countries show higher involvement due to their greater industry orientation and experience in collaborative research (Nokkala, Heller-Schuh and Paier 2011). In a dynamic perspective, literature detected that the programmes participation path has led to the creation of a group of persistent "happy few", whose level of participation is very high and stable over time, while remaining organizations participate occasionally or do not participate at all. It is thus interesting to investigate about the mobility between these groups. Our research hypothesis mainly rely on the idea that a change – positive or negative - in the degree of participation at individual level could be affected by a series of characteristics/factors that could be roughly classified as follow: a) individual characteristics such as size, reputation, subject composition; b) national economics; c) differences in national research policies and strategies; d) differences in the FPs composition.

Methodology and data sources

Data sources

The paper is based on an extended version of the European Tertiary Education Register database (ETER; https://www.eter-project.com/). ETER is the reference dataset on European higher

education. It provides a comprehensive coverage of European HEIs graduating at least at the bachelor level and includes a wealth of statistical data including institutional characteristics, revenues and expenditures, students and graduates. The extended version of ETER that has been developed within the RISIS project (risis.eu) includes data on scientific publications from the Leiden ranking database and the number of participation in EU-FP projects. The dataset also include 2008 data with the addition of EUMIDA information. Data include information on 28 countries for 2 years, 2008 and 2015, with 1,239 HEIs and 2,478 observation in total. The sample provides a good coverage of Europe, including some Nordic European countries (DK, FI, LT, LV, NO, SE), Western Europe countries (AT, BE-Flanders, CH, DE, IE, LU, NL and UK), Southern Europe countries (CY, ES, GR, IT, MT and PT), as well as a number of Eastern European countries (BG, CZ, EE, HU, PL, RO, SI and SK).

Variables

Variables included in the current step of the analysis are listed below. For all of them, absolute values and the difference between 2015 and 2008 have been calculated.

- *Number of EU-FP participations*: number of participation for each institution for starting year. In a further step, a network centrality measure will be included (we expect this measure to be strongly correlated with the number of participations).
- *Size measure*: we use the total academic staff in FTE (Full Time Equivalent).
- *Education intensity* as the ratio between the total number of undergraduate students and the total staff in FTE.
- *Research volume*: approximated by the total students enrolled in a doctoral course normalized by the maximum in our sample.
- *Subject specialization*, measured by the Herfindahl Index of the distribution of the doctoral students by the ten fields of educational statistics.
- *National economics* (GPD per capita in PPP, research and development expenditure by sector)
- *Information about the composition of different FPs.*

Methods

As a first step, we provide descriptive analysis of changes between the two years by grouping the HEIs in our sample in classes according to the number of participations.

Classes are defined as follows:

- Class 0: no participation
- Class 1: from 1 to 4 participation
- Class 2: from 5 to 14 participation Class 3: from 15 to 40 participation
- Class 4: more than 40 participation.

Class ranges has been defined according to the number of participation corresponding to precise percentile of the sample (50%, 75%, 90%, 95%, 99%) according to the statistics of 2015. On this basis, we firstly build up a transition probability matrix, highlighting the probability of transitioning from one state to another:

$$p_{ij} = Pr\{X_t = j | X_{t-1} = i\} \qquad (1)$$

A transition probability is the probability of being in state j at the end of the branch given that the process was in state "*i*" at the start of the branch. In our model, "*t*" is 2015 and "*t-1*" is 2008. "*i*" and "*j*" represent different classes of participations corresponding to different groups in the selected years. We consider transitions between the year 2008 and 2015, a time period which

includes four different FPs (from FP5 to Horizon 2020) and is sufficiently long to observe a relevant, if present, mobility.
Once defined the above-mentioned probabilities, we investigate the variables that affect probability for an HEI to move between classes, upgrading or downgrading the intensity of participation at individual level. To do this, we group the in and out mobility from 2008 to 2015 as follows: "group -1" includes HEIs that move from an higher to a lower class, "group 0" HEIs in the same class both in 2008 and 2015 and "group 1" HEIs that move from a lower to an higher class.
As further steps of the study we:

- will further refine the class identification with the help of the "*k-means*", a method of vector quantization used in the cluster analysis (to identify "natural breaks" in the distribution);
- will more accurately define some variables such as , especially the one related to the research volume and the number of participation, testing also a measure of network centrality;
- will investigate on factors and variables that affect the transition probabilities between status (in terms of class of participation) mainly with a fractional logit model (Papke and Wooldridge 1996), that permits the analysis of dependent variables with fractional or proportional values in the unit interval, i.e., $y \in [0,1]$.

- $E(y|x) = exp(x\beta)/(1 + exp(x\beta))$

As a second step, we will regress relevant individual HEIs' characteristics and external variables on the differences in HEIs degree of participation with the aim to detect variables that affect the probability of an in or out mobility from the "happy few" network. The most appropriate approach seems to be the estimation with first differences. The base case would be

$$y_{it} = c_i + x_{it'}\beta + \varepsilon_{it} \qquad (2)$$

Which implies the first differences equation

$$\Delta y_{it} = \Delta c_i + (\Delta x_{it})'\beta + \Delta\varepsilon_{it} \qquad (3)$$

Preliminary results

Figure 1 displays the transition matrix for the period 2008-2015. Within this matrix, in 2008 the HEI "*i*" is in a specific state (class) x_i and at the time *i+1* the same HEI changes to state (class) x_i+1 according to the given transition probabilities.

		2015					
		class					
	class	0	1	2	3	4	Total
2008	0	591	87	3	0	0	681
		86.78	**12.78**	**0.44**	**0**	**0**	**100**
	1	77	156	33	0	0	266
		28.95	**58.65**	**12.41**	**0**	**0**	**100**

		2	**46**	**72**	**22**	**0**	**142**
	2	**1.41**	**32.39**	**50.7**	**15.49**	**0**	**100**
		0	0	31	80	11	122
	3	**0**	**0**	**25.41**	**65.57**	**9.02**	**100**
		0	0	0	9	19	28
	4	**0**	**0**	**0**	**32.14**	**67.86**	**100**
	Total	670	289	139	111	30	1,239

Figure 1. Transition probabilities matrix

Numbers in first row of each class refer to the number of HEIs that belong to a selected class in 2008 and in 2015, while the second row refers to the corresponding transition probabilities. For example, the probability that a HEIs belong to class 0 both in 2008 and 2015 is 86.78%, while the same probability for the same HEIs to shifts from class 0 in 2008 to class 1 in 2015 is 12.78% while from zero to 3 is even smaller (0.44%).

As expected, the matrix displays a good deal of stability. Overall, ¾ of HEIs in our sample stayed in the same class in both periods. At the same time, a certain mobility between classes emerges. Compared to 2008, in 2015 90 HEIs shift their status from not participant to participant, and 66 (33 from class 1 to class 2, 22 from class 2 to class 3 and 11 from class 3 to class 4) will register a positive shifting. On the other hand, 165 HEIs moved from a higher to a lower class. Looking at the transition probabilities, it is interesting to note that the probability of an out-mobility with a downgrade in previous classes is higher than the corresponding probability of an upgrade (due to a rise in the number of participations). For example, the probability that an HEI belonging to class 2 in 2008 will transit to class 1 in 2015 is 32.4 % while the same figure for a shift from class 2 to class 3 is about 15.5%. Overall, the core seems to have become slightly smaller over time, even if the differences is limited. Interestingly, mobility between the participant and non-participant group is more limited than vertical mobility, suggesting a strong entry barrier for non-participating HEIs.

Looking to countries, a certain pattern seems to emerge with some countries displaying clear patterns (upward or downward), others. To take into account the scale of status variation, we decide to compare countries through a mobility index. For each country, that index will weigh the magnitude of the status variation of HEIs: if an HEI shifts its status of two classes then is weight will be -2 while for the opposite the weight is +2 and so on. Figure 2 illustrates the results.

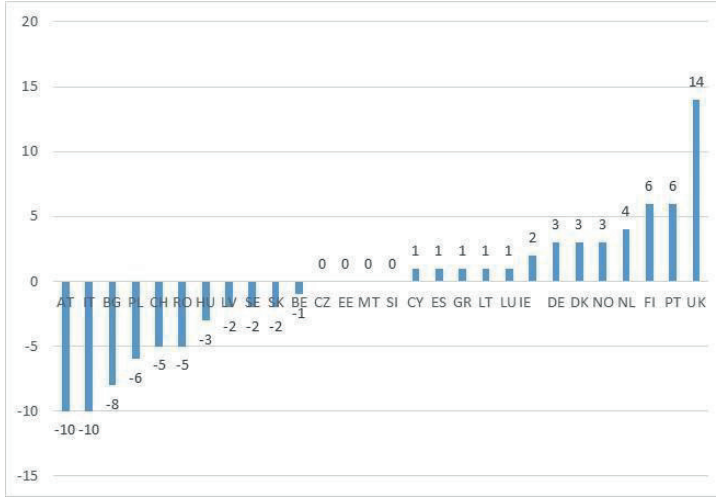


Figure 2. Weighted in and out mobility index

Mobility index scores evidence that countries with significant out mobility are Austria, Bulgaria, Switzerland, Italy, Poland and Romania. For Switzerland, this downgrade is mainly referable to the fact that the country was not fully associates from the beginning to H2020, causing a significant drop of participations in the recent year. Within this group, we notice the presence of three east European countries: BG, PL and RO, a worrying finding in terms of cohesion in the European Research Area. A more precise picture of the mobility between classes and groups claims for the analysis of 2008-2015 differences for size, education intensity, research volume and subject concentration variables; we then compare the median values of changes of these variables between 2008 and 2015 between three groups of HEIs: those that did not change of class (0), those that moved to a lower class (-1) and those that moved higher (1). This comparison could permit to capture differences in the main characteristic shaping the HEIs. This comparison helps to develop a more accurate analysis: for example, institutions included in class zero in 2008 that did not changes their status in 2015 might differ by some characteristics from those that shifted to group 1. We then performed a k-sample test on the equality of the medians, for absolute values and differences between 2008 and 2015, with the reference to the median of HEIs not changing their class. Figure 3 summarizes results for two variables that are likely to be associated with mobility, i.e. the size in terms of academic staff and the research volume.

	Academic staff FTE			*Difference 2008-2015*		
Class	*-1*	*0*	*+1*	*-1*	*0*	*+1*
0		133	322***		19.7	37.93
1	402***	645.5	846	42.56	37.91	194.14***
2	794.3***	1247.5	1153.4	136.14	248.85	142.52
3	1871	1878	2733.78***	317.50	346.02	515.50***
4	2431	2927.3		453.78	581.00	

	Research volume			*Difference 2008-2015*		
Class	*-1*	*0*	*+1*	*-1*	*0*	*+1*

0		0	.00423***		0.0000	0.0004***
1	.00876***	.02793	.03534	0.0005***	0.0015	0.0057
2	.05406***	.06765	.10210***	0.0051***	0.0040	0.0116***
3	.13321	.13525	.20550	0.0028	0.0206	0.0586
4	.17662	.21446		0.0438	0.0424	

Figure 3. Median of the difference 2008-2015 for few selected variables

Figures point out that:

a) HEIs that shift their status from not participant to participant: these HEIs had already in 2008 a larger size and higher research volume that those that continued not to participate and had a greater increase as well (difference is however non-significant for staff). This might be interpreted as 'screaming out' the large and more research oriented non participating HEIs. The opposite applies to HEIs that ceased to participate.

b) In nearly all cases, HEIs that shift their status to a more intense participation (group 1) have usually larger differences in median values that HEIs that shift to the lower one (group 1) or are invariant. This applies both to 2008 values and to changes over the considered period. The opposite applies to HEIs moving downward.

We have to keep in mind that data for shifting HEIs for class 3 and 4 are based on a very small number of observations – due to the limited number of HEIs interested. This of course affects statistical significant.

Discussion and further work

While most of the literature focused either on stability in participation networks, or on factors affecting the participation level in a static (cross-sectional) perspective, this papers aims to shed light on the extent of change over time and on its determinants. Overall, we could show some degree of change in participation over the considered period; a more fine-grained analysis suggested that, while there is closure in the group of participant HEIs, level of participations might vary significantly over time. In a further step, time-series analysis might allow a more precise distinction between transient participants and regular participants. Preliminary results also display meaningful patterns in terms of starting conditions and of change over time in HEI characteristics for those HEIs that changed significantly their participation level. As expect, size and research volume seem to have a direct impact on mobility.

Moreover, the analysis displays a clear country pattern in participation mobility, with several selected countries, partially located in the Eastern Europe, more interested by an out mobility. In general, the probability of a status downgrade is higher than the corresponding upgrading one. In status shifting, size and research volumes seems directly matter: their rise will probably positive impact on mobility. Focusing on median values, upgrading HEIs have higher figures, while for downgrading ones groups' medians appear to be closer; this seems to suggest that there could be different variables, possibly exogenous, that could affect their mobility.

As highlighted in the methods section, more refined statistical methods will allow in a further step a more precise identification of the respective role of different factors in participation change.

References

Barber, M. J., Krueger, A., Krueger, T. & Roediger-Schluga, T. (2006). Network of European Union–funded collaborative research and development projects. *Physical Review E, 73(3)*, 036132.

Enger, S. G. & Castellacci, F. (2016). Who gets Horizon 2020 research grants? Propensity to apply and probability to succeed in a two-step analysis. *Scientometrics, 109(3)*, 1611-1638.

Geuna, A. (1998). Determinants of university participation in EU-funded R & D cooperative projects. *Research Policy, 26*, 677-687.
Heller-Schuh, B., Barber, M., Henriques, L., Paier, M., Pontikakis, D., Scherngell, T., Veltri, G. A. & Weber, M. (2011). *Analysis of Networks in European Framework Programmes (1984-2006)* Luxembourg: Publications Office of the European Union.
Henriques, L., Schoen, A. & Pontikakis, D. (2009). Europe's top research universities in FP6: scope and drivers of participation. *JRC Technical Notes, 53681*.
Lepori, B., Heller-Schuh, Scherngell & Barber (2014). Understanding Factors Influencing Participation to European Programs of Higher Education Institutions. *Sti 2014, Leiden*.
Luukkonen, T. (2013). The European Research Council and the European research funding landscape. *Science and Public Policy, 41(1)*, 29-43.
Luukkonen, T. & Nedeva, M. (2010). Towards understanding integration in research and research policy. *Research Policy, 39(5)*, 674-686.
Mataković, H. & Novak, I. R. (2013). Croatia's participation in the Seventh Framework Programme: a Moderate Success? *Business Systems Research, 4(2)*, 126-143.
Merton, R. K. (1988). The Matthew Effect in Science, II. Cumulative Advantage and the Symbolism of Intellectual Property. *Isis, 79*, 606-623.
Nedeva, M. (2013). Between the global and the national: Organising European science. *Research Policy, 42(1)*, 220-230.
Nokkala, T., Heller-Schuh, B. & Paier, M. (2011). Ranking lists and european framework programmes. In Anonymous *Public Vices, Private Virtues?* (pp. 111-139). Springer.
Papke, L. E. & Wooldridge, J. M. (1996). Econometric Methods for Fractional Response Variables with an Application to 401 (K) Plan Participation Rates. *Journal of Applied Econometrics, 11(6)*, 619-632.
Primeri, E. & Reale, E. (2012). How Europe shapes academic research: Insights from participation in European union framework programmes. *European Journal of Education, 47(1)*, 104-121.
Scherngell, T. & Barber, M. (2011). Distinct spatial characteristics of industrial and public research collaborations: evidence from the fifth EU Framework Programme. *The Annals of Regional Science, 46(2)*, 247-266.
Seeber, M., Lepori, B., Montauti, M., Enders, J., De Boer, H., Weyer, E., Bleiklie, I., Hope, K., Michelsen, S. & Mathisen, G. N. (2015). European universities as complete organizations? Understanding identity, hierarchy and rationality in public organizations. *Public Management Review, 17(10)*, 1444-1474.

From closed to open access: A case study of flipped journals

Fakhri Momeni[1], Philipp Mayr[1], Nicholas Fraser[2] and Isabella Peters[3]

[1] *firstname.lastname@gesis.org*
GESIS – Leibniz Institute for the Social Sciences, Unter Sachsenhausen 6-8, 50667 Cologne (Germany)

[2] *N.Fraser@zbw.eu,* [3]*I.Peters@zbw.eu*
ZBW – Leibniz Information Centre for Economics, Düsternbrooker Weg 120, 24105 Kiel (Germany)

Abstract

In recent years, increased stakeholder pressure to transition research to Open Access has led to many journals 'flipping' from a toll access to an open access publishing model. Changing the publishing model can influence the decision of authors to submit their papers to a journal, and increased article accessibility may influence citation behaviour. The aim of this paper is to show changes in the number of published articles and citations after the flipping of a journal. We analysed a set of 171 journals in the Web of Science (WoS) which flipped to open access. In addition to comparing the number of articles, average relative citation (ARC) and normalized impact factor (IF) are applied, respectively, as bibliometric indicators at the article and journal level, to trace the transformation of flipped journals covered. Our results show that flipping mostly has had positive effects on journal's IF. But it has had no obvious citation advantage for the articles. We also see a decline in the number of published articles after flipping. We can conclude that flipping to open access can improve the performance of journals, despite decreasing the tendency of authors to submit their articles and no better citation advantages for articles.

Introduction

The term "flip" is often used to describe a journal transitioning from a subscription-based (or closed access (CA)) to an open access (OA) publishing model (Solomon, Laakso & Björk, 2016). This change in the publishing model might have consequences for the journal's future development, for example in terms of citation rates or publishing volumes. Most evidence points towards an open access citation advantage for OA articles over CA articles overall (OACA[1]; Piwowar & Vision, 2013; Sotudeh et al., 2015; Swan, 2010), although the citation advantage is found to be driven primarily by 'Green' and 'Hybrid' OA whilst articles published in 'Gold' OA journals conversely show a disadvantage in terms of citations versus CA articles (Archambault et al., 2014; Piwowar et al., 2018). Swan (2010) reported the results of studies on the OACA across multiple fields and showed significant variability between fields. McKiernan et al. (2016) compared the relative citation rate of OA articles for 19 fields and showed the highest citation advantage for agricultural studies, physics/astronomy and medicine, but a citation disadvantage for ecology and communications studies. A randomized control trial conducted by Davis (2011) could also not confirm the citation advantage of OA journals over CA journals. Hence, there remain many unanswered questions in what drives this erratic behaviour. In addition, there are few studies that systematically research the effects flipping has on a journal's standing, in terms of its citation rates and impact factor (IF) as well as its publication output. One exception is a study by Busch (2014a) who demonstrated an increase in IFs for five journals published by BioMed Central after flipping to an OA model. But, he also found a reduction in the number of articles published by four journals of this group (Busch 2014b).

An important concern for journals is to find an alternative source of income to reader subscription charges when flipping to OA. One alternative source is the author-pays model that shifts the publication costs to authors or their institutions. This may generate financial barriers for some authors who wish to submit their articles to such journals. According to

[1] The Open Access Citation Advantage Service (OACA) list [Internet]. Sparceurope.org. 2017 [accessed 27 January 2019]. Available from: https://sparceurope.org/what-we-do/open-access/sparc-europe-open-access-resources/open-access-citation-advantage-service-oaca/

Solomon, Laakso & Björk (2016), transitioning a journal to an OA model for those societies with low numbers of publications can be expensive. Björk (2012) additionally demonstrated that the author-pays model in hybrid journals is unpopular with authors. Peterson et al. (2013) argued that the cost of article processing charges (APC) in this model is often too much for many authors, and publishers try to solve this problem in different ways such as fee waiver policies, subsidizing academic publishing directly without profiteering intermediaries, etc.
Such consequences are of immense importance to those responsible for journal management, thus in-depth, longitudinal bibliometric studies can help to inform decision making of publishers and societies, and their assessment of chances and risks of flipping their journals. A recent study on "reverse-flipped" journals by Matthias et al. (2019) shows that a majority of these journals had an experience of flipping to OA before, pointing out that flipping was not successful for them for different reasons. Moreover, authors need to know what to expect from flipped journals, and whether the move from CA to OA negatively or positively affects the quality of the journal, its reputation, or the chances to get articles published.

This is a work in progress paper presenting the first results of a study on the evolution of bibliometric indicators for a sample of journals, over a period of several years before and after the journal flipped to an OA publishing model. We assess the effect of the journals flipping to OA by studying the changes in the number of published articles, as well as changes in citation metrics at the article (average relative citations) and journal (IF) level[2].

Data and methods

A list of journals that have flipped to OA, and the years in which they flipped were retrieved from the Open Access Directory (a wiki where the OA community can create and support simple factual lists about open access to science and scholarship)[3], provided by the School of Library and Information Science at Simmons College. From this list we retrieved details of 235 journals that flipped to OA prior to 2017[4]. Of these journals, 171 could be matched to journals indexed in the Web of Science (WoS) via matching of names and ISSNs.
The effect of flipping to OA will be measured by a descriptive analysis of the number of articles published by the flipped journals, the number of citations to those articles (using the matching citation algorithm implemented by the German Competence Centre Bibliometrics: http://www.forschungsinfo.de/Bibliometrie/en/index.php) and the respective journal IFs. Additionally, we will give an overview on when the most journals converted to OA.
We calculated the two years IF following the same methodology employed by Clarivate Analytics[5] for each of the 171 journals for each year. Based on this definition, the IF is defined as all citations to the journal in the current year to items published in the previous two years, divided by the total number of scholarly items (these comprise articles, reviews, and proceedings papers) published in the journal in the previous two years. In order to be able to compare IFs between different fields, they were field-normalized using the rescaling method introduced by Radicchi, Fortunato & Castellano (2008), in which the citation rate calculated in the IF definition is rescaled by dividing by the arithmetic mean of the citation rate of all articles in its discipline. The 252 subject categories included in WoS were applied for the field-normalization. In this classification system, journals can have more than one category, therefore we considered the mean citation rate of all articles in all categories of which this journal belongs to.
To study the evolution of citations at the article level, we calculate a relative citation (RC) count for each individual paper published within the journals in our dataset, normalised to

[2] http://www.science-metrix.com/?q=en/expertise/bibliometrics/methods
[3] http://oad.simmons.edu/oadwiki/Journals_that_converted_from_TA_to_OA
[4] See the list of journals here: https://github.com/momenifi/OASE
[5] https://clarivate.com/essays/impact-factor/metric

account for different citation patterns across fields. The RC of a paper is calculated for each year by computing the sum of citations gained by the individual paper, divided by the average number of citations of all papers across its field(s) published in the same year. The window to compute the number of citations is three years (to be sure the articles receive their citation peak). An RC value above 1 means that a paper is cited more frequently than the average citation level for all papers in that field, and vice versa. To compare the citations for two groups of papers, we therefore calculate the mean RC of all papers for each group, referred to as the average of relative citations (ARC). In this way we can compare the ARC of multiple groups of papers across different years and different fields.
To be able to compare changes between our dataset of OA journals and a dataset of CA journals (obtained from the ScienceDirect website[6]) across multiple scientific domains we reclassified our dataset of journals into the six major domains of the Science Metrix classification[7] system described by Archambault et al. (2011). This classification system allows us to make coarser comparisons between major scientific domains in comparison to than the 252 subject categories in WoS.

Journals that had no flipping date indicated in the Open Access Directory were excluded from our analysis. Hence, we present results with a number of journals (n=171) that will be indicated in the Figure 1. Also, for our analyses of article published volumes, ARC and IF, we excluded the journals that had no articles, ARC and IF for at least one year of the considered years.

Results

IFs are calculated based on citations earned by articles published in the two past years, thus we expect to observe the impact of converting to OA at least one year after the flip. Due to the journal review time – e.g. when a journal flips, newly submitted articles will take several months to proceed through the review process. Therefore we considered the following year as the flipping point for journals which were flipped in the fourth quarter to ensure that articles reflect the OA model under which they were submitted. Figure 1 shows the distribution of years in which journals in our WoS dataset flipped. We observe a peak in the number of flipped journals in 2006, as well as a long-term steady increase in the number of journals that have converted to OA across all years. The peak in 2006 is caused by a large number of journal conversions carried out by two major publishers: Hindawi and the Spanish National Research Council. According to the Open Access Directory, no journals flipped in 2016.

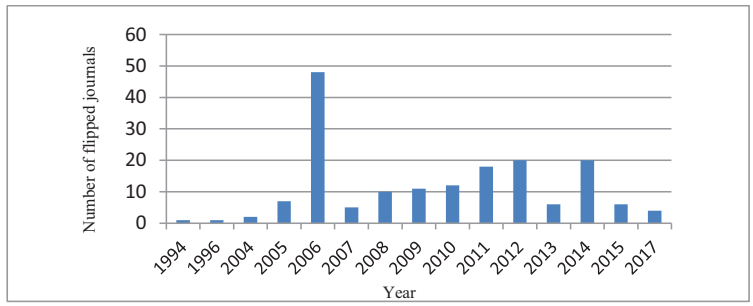


Figure 1. Distribution of flipping years for 171 WoS journals.

After a journal flips to OA, the costs of publishing transform from the reader side (i.e. subscriptions) and must be recovered from alternative sources, commonly in the form of

[6] https://www.sciencedirect.com/browse/journals-and-books?accessType=subscribed
[7] http://science-metrix.com/?q=en/classification

Articles Processing Charges (APC) paid by funders, institutes or authors directly (the 'Gold' OA model). Authors or institutes without access to sufficient funding sources may therefore be unable to publish in converted OA journals, which may influence the number of submitted articles. Conversely, some journals may flip to OA with the financial support of a funding body or society and not require publishing fees on the side of the author (the 'Diamond' OA model), which may influence published article volumes differently. Figure 2 shows the average number of articles published by journals four years before and after flipping. After excluding the journals that have at least one year with zero published articles and flipped after 2012, we obtain a dataset of 58 journals. From this figure, we observe a rising trend in the average number of publishing articles in the years preceding flipping when considering all journals (thick red line), followed by a decline in the number of published articles in the year after flipping (from an average of 142 published articles one year before flipping, to 132 one year after flipping). Two years subsequent to the journal flipping, published article volumes become more stable. When considering individual flipping years, we also observe a decrease in the number of articles published after flipping, with the sole exception of journals that flipped in 2009. Overall, these findings agree with those from Björk (2012) and others, that hybrid journals publish more articles in CA than OA.

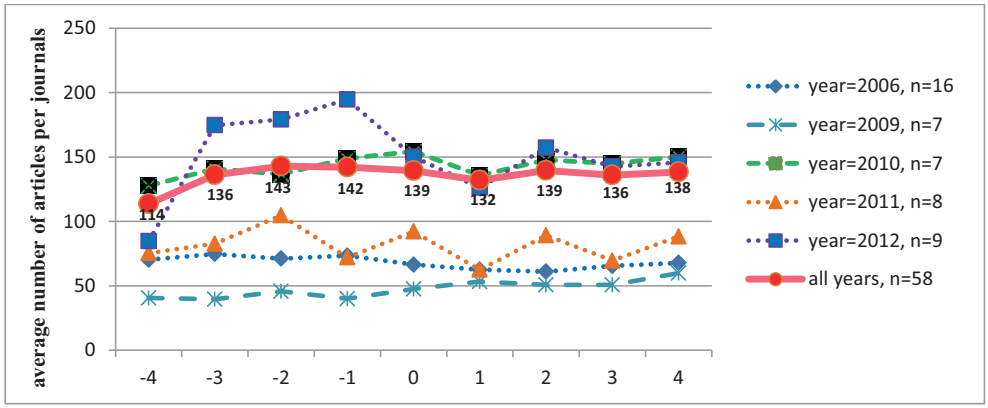


Figure 2. Yearly average number of articles for flipped journals four years before and after flipping (n=number of journals and year is the flipping time). Read 0 on the x-axis as the year of flipping; -1 is the year before and 1 the year after flipping.

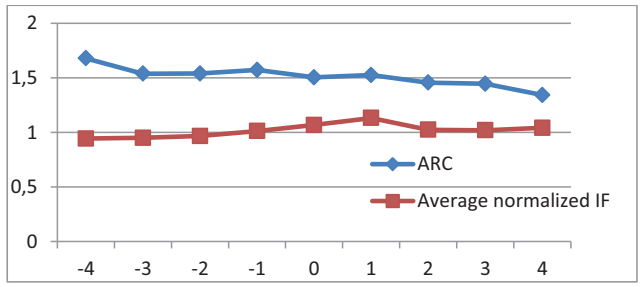
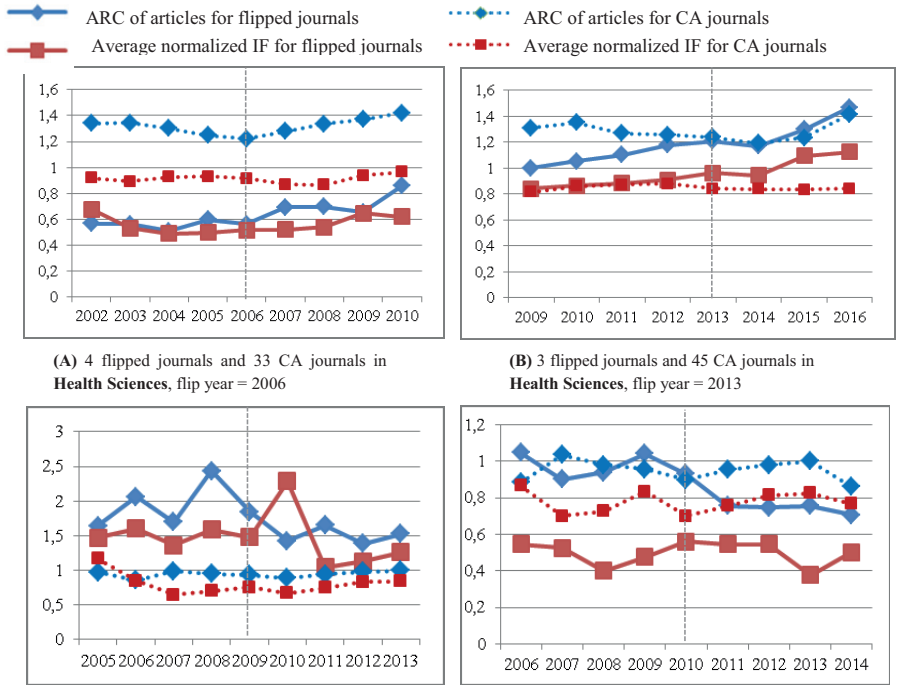


Figure 3. Comparing the ARC of articles and average normalized IF of 43 flipped journals before and after flipping.

Figure 3 shows the IF and ARC for journals and articles, respectively, in our dataset for the four years before and after flipping. We only included journals with IF and at least one article

with and RC larger than zero for each of the years analysed and flipped until 2012. We observe a decrease in ARC after flipping, indicating a lower impact for articles published in these journals after flipping. However, the decrease in ARC occurs at the same time as a peak for IF one year after flipping. Our interpretation for this peak is that when the journals flip they probably also make the old articles OA, and so the increase in IF would be due to increased citations to articles that were published under the CA model, but have now been made OA. In general, we observe higher IFs in the years after flipping than before.

Flipping to OA can affect ARC and IF differently across different fields. To see the difference, we categorised the journals based on flipping years and fields, and compared their ARC and normalized IF with CA journals in the same field. Figure 4 shows a sample of the fields with the highest numbers of journals in each category. We observe that flipped journals experience more strongly fluctuating long-term trends of ARC and IF across the time period of our measurement, in comparison to those of the CA journals. However, for Health Sciences (A and B) we see an improvement in the ARC and IF after flipping, but journals in Applied Sciences (C and D) show no long-term advantage for these indicators for both years 2009 and 2010. However, we note that there are limited numbers of flipped journals across our samples in this analysis, and thus analysing these relationships more closely will be part of the focus of our future.



(A) 4 flipped journals and 33 CA journals in **Health Sciences**, flip year = 2006

(B) 3 flipped journals and 45 CA journals in **Health Sciences**, flip year = 2013

(C) 4 flipped journals and 14 CA journals in **Applied Sciences**, flip year = 2009

(D) 3 flipped journals and 14 CA journals in **Applied Sciences**, flip year = 2010

Figure 4. Comparing ARC of articles and normalized IF of flipped and CA journals across some fields.

Conclusion and future work

The literature reporting studies on flipped journals shows that journals IF usually increase after flipping (Busch, 2014a). Our results agree with these previous findings, but show that

whilst IF increases in the years immediately following flipping, the article-level ARC decreases, in line with results found by Piwowar et al. (2018) and Archambault et al. (2014) that gold OA articles are cited less than CA articles. We also observe that a lower number of articles are published after flipping, pointing to a lower tendency amongst authors to submit to OA journals. So far, this work in progress studied only a limited number of flipped journals in the WoS with different flipping years and in different fields. Hence, future work will expand this study to Scopus to consider a higher number of flipped journals, and to research in greater detail the publication output of those journals before and after flipping, e.g. in terms of article numbers or length. We will also study in more detail the so-called "quality effect", which assumes that high quality publications profit proportionally more from OA, as they are more citable than low-quality publications (Ottaviani, 2016). The question we seek to answer is whether journals with high IFs benefit significantly more from flipping to OA than journals with medium or low IFs. Our bibliometric study will be complemented with more qualitative information from interviews with authors, editors, and publishers to reveal their attitudes towards journal flipping and OA, their expectations regarding journal quality and indicators as well as their motivation to change the publication model. With this we hope to better show such interwoven factors that influence bibliometric indicators of flipped journals.

Acknowledgement

This work is supported by BMBF project OASE, grant number 01PU17005A.

References

Archambault, É., et al. (2014). Proportion of open access papers published in peer-reviewed journals at the European and world levels – 1996-2013. https://digitalcommons.unl.edu/scholcom/8/

Archambault, É., Beauchesne, O. H., & Caruso, J. (2011). Towards a multilingual, comprehensive and open scientific journal ontology. In Proc. of the 13th ISSI conf. (pp. 66-77). South Africa.

Björk, B. C. (2012). The hybrid model for open access publication of scholarly articles: A failed experiment?. JASIST, 63(8), 1496-1504.

Busch, S. (2014a). "The careers of converts – how a transfer to BioMed Central affects the Impact Factors of established journals" BioMed Central (blog).

Busch, S. (2014b). "The Impact Factor of Journals Converting from Subscription to Open Access." BioMed Central (blog).

Davis, P. M. (2011). Open access, readership, citations: a randomized controlled trial of scientific journal publishing. The FASEB Journal, 25(7), 2129-2134.

Matthias, L., Jahn, N., & Laakso, M. (2019). The Two-Way Street of Open Access Journal Publishing: Flip It and Reverse It. Publications, 7(2), 23.

McKiernan, E. C., et al. (2016). How open science helps researchers succeed. eLife, 5, e16800. https://doi.org/10.7554/eLife.16800.001

Ottaviani, J. (2016). The post-embargo open access citation advantage: it exists (probably), it's modest (usually), and the rich get richer (of course). PloS one, 11(8), e0159614.

Peterson, A. T., et al. (2013). Open Access and the author-pays problem: assuring access for readers and authors in a global community of scholars. https://doi.org/10.7710/2162-3309.1064

Piwowar, H., et al. (2018). The State of OA: A large-scale analysis of the prevalence and impact of Open Access articles. PeerJ, 6, e4375. https://doi.org/10.7717/peerj.4375

Piwowar, H. A., & Vision, T. J. (2013). Data reuse and the open data citation advantage. PeerJ, 1, e175. https://doi.org/10.7717/peerj.175

Radicchi, F., Fortunato, S., & Castellano, C. (2008). Universality of citation distributions: Toward an objective measure of scientific impact. Proceedings of the National Academy of Sciences, 105(45), 17268-17272.

Solomon, D. J., Laakso, M., & Björk, B. C. Converting scholarly journals to open access: a review of approaches and experiences [Internet]. Harvard Library Office for Scholarly Communication; 2016

Sotudeh, H., Ghasempour, Z., & Yaghtin, M. (2015). The citation advantage of author-pays model: the case of Springer and Elsevier OA journals. Scientometrics, 104(2), 581-608.

Swan, A. (2010). The Open Access citation advantage: Studies and results to date.

Reliability of Scopus author identifiers (AUIDs) for research evaluation purposes at different scales

David Campbell[1] and Brooke Struck[2]

[1] *david.campbell@science-metrix.com*
Science-Metrix Inc., 1335 Mont-Royal est, Montréal, QC (Canada) H2J 1Y6

[2] *brooke.struck@science-metrix.com*
Science-Metrix Inc., 1335 Mont-Royal est, Montréal, QC (Canada) H2J 1Y6

Abstract

This study investigates the reliability of the Scopus author identifiers (AUIDs) for research evaluation purposes at different scales. Differences in the scientific performance of researchers across funding bins were first computed using manually cleaned publication portfolios for populations of various sizes (from 20 up to 1,500 researchers per bin, for 5 bins; i.e., for populations from 100 to 7,500 researchers). The differences in performance across bins based on the AUIDs were then compared to those based on the cleaned portfolios (the "ground truth"). To the degree that the findings based on the AUIDs approximate those based on the manual cleaning, the AUIDs can reliably be used in place of manual cleaning. The results show that when comparing groups of fewer than 500 researchers, manual cleaning should be used. With groups of roughly 500 researchers, manual cleaning should still proceed if budget is sufficient. Otherwise, AUIDs can be used providing that care is taken not to overinterpret the findings. With groups of around 1,000 or more researchers, the use of Scopus AUIDs is warranted to decrease study costs. At that scale, the AUIDs provide highly reliable conclusions in terms of direction and size of effect.

Introduction

In evaluating the benefits of a funding program, researcher performance data are always at a premium, whether survey data, interview data, case studies or data of other kinds. Bibliometric databases offer a plentiful source of information and provide a basis for large-scale, powerful studies. However, significant efforts are often required to disambiguate authorship data: connecting publication outputs to the correct researchers.

High-quality data are crucial to ensure buy-in in the study findings as well as the credibility and transparency of bibliometric analyses. Accordingly, Science-Metrix has always performed manual disambiguation, aided by algorithms, of the researchers' publication data to achieve very high precision and recall. However, manually cleaning data for a large number of researchers (several thousands), as in national assessments, is not cost-effective.

In recent years, several algorithms producing unique author identifiers have emerged to address the challenge of disambiguating researcher data (D'Angelo, Giuffrida, & Abramo, 2011; Gurney, Horlings, & van den Besselaar, 2012; Tang & Walsh, 2010). One such example is the Scopus author identifier (AUID) by Elsevier. However, Science-Metrix' working assumption has been that these tools—useful as they may be for other purposes—were not yet sufficiently developed to underpin robust performance measurement in an evaluation context.

The objective of this research paper is to test that assumption at different scales (i.e., for populations of different sizes). The analyses presented here will provide a quantitative assessment of the reliability of one such tool (the Scopus AUID) for research evaluation. This assessment will go beyond the analysis of the AUIDs' recall and precision alone, as is usually done in studies aiming to develop author disambiguation algorithms. Additionally, this study will measure differences in the scientific performance of groups of researchers obtained using the AUIDs, comparing these to differences that one would achieve based on manual data disambiguation. A similar approach was used by Strotmann & Zhao (2012), but rather than comparing manual to automated author disambiguation, they compared automated author disambiguation to a simple approach based on author surnames and first initials. Nevertheless,

their focus on assessing quality relative to a practical use case (rather than just precision and recall) is valuable, and we follow that same path here in our work.

Over the past 15 years, Science-Metrix has evaluated many programs for national funding agencies and for public/private foundations around the world (Picard-Aitken et al., 2010; Science-Metrix, 2018). A 2013 project involved the manual cleaning of over 10,000 researcher portfolios. The performance of those researchers was compared across funding bins to assess the added value of the support they received from a funder. Two key dimensions of relevance were publication output volume and citation impact. The 10,000 manually cleaned portfolios provide the "ground truth"—acknowledging that even manual cleaning is imperfect—against which the reliability of the AUID is benchmarked here. The approach to compare performance across funding bins is deployed twice, once based on the manual cleaning and once based on the AUID, and then the resulting findings are compared to each other. To the degree that the findings based on AUID approximate those based on the manual cleaning, the AUID can reliably be used in place of manual cleaning.

Methods

The study was conducted on the Scopus database, produced by Elsevier. Science-Metrix' local instance of Scopus includes considerable data enrichment, including the legacy of institutional and researcher profiles manually cleaned through the history of studies performed at the firm.

From the 2013 project noted above, some 10,000 researcher portfolios were manually cleaned. These researchers were those who applied for funding during competition years 2009–2013. Their papers were retrieved from publication years 2000–2011; as publication years 2012 and 2013 were not complete in the database at the time, these were excluded from the analysis, both for the original project and in the present case. Each researcher's performance was assessed for the five years prior to their grant application.

Each researcher was linked to one and only one AUID. For the present study, their "main" AUID was identified as the one in their cleaned profile that had the largest number of papers associated with it. This approach was easily automated in the present context, with the benefit of the manually cleaned data already in hand, and thus did not represent a substantial investment of additional time. However, in the context of a real program evaluation, sufficient time must be allocated to this step; using the AUID dramatically reduces the costs of data preparation, but there are still some time costs involved.

Given its previous experience working with AUIDs, Science-Metrix also proactively excluded the most problematic AUIDs, those that were most evidently spurious. Any AUID was excluded if its publications were spread across more than 34 subfields (of 176) or 10 fields (of 22) in the Science-Metrix taxonomy of research (Archambault, Beauchesne, & Caruso, 2011).

There was one further exclusion criterion. Some AUIDs are associated with multiple surnames, which is important to account for misspellings, inversions of compound names, and so forth. However, one would expect relatively little variation in the first letter of those surnames; for example, a compound name like "van den Berg" might be indexed in several different ways, with the first letters of these variations being "v," "d," or "b." Problematic cases were conservatively identified as single AUIDs associated with 10 or more letters as the first letter in the various permutations of surname, and these cases were removed.

In total, these filters removed 0.2% of the study population. While small in number, these problematic AUIDs are often associated with huge numbers of publications. Their removal is thus crucial, as such outliers could substantially influence the results obtained. Further problematic AUIDs were identified during validation but could not be removed by automated means without also removing many good AUIDs. Given that there were only 22 problematic

instances, it is highlighted here that manual cleaning could be used in such cases to fill in the gap and ensure high-quality data.

The test population presented here is made up of North American researchers active in the natural and applied sciences. Since the Scopus AUIDs are more problematic with, for example, Asian names, the share of faulty AUIDs might be higher in other contexts where the above filters might not work. More research on these screening criteria across geographic and thematic areas is needed to make them effective in the widest range of cases.

Within the pool of 10,000 cleaned portfolios, study populations of varying sizes were created using random sampling. Tests were conducted using populations of the following numbers of researchers: 100, 500, 1,000, 2,500, 5,000 and 7,500. Researchers were labelled by their peer-review score as well as their funding bin, of which there were five: **Bin 0:** unfunded, **Bin 1:** lowest funding tier, **Bin 2:** 2nd funding tier, **Bin 3:** 3rd funding tier, and **Bin 4:** highest funding tier. The distribution of researchers along the peer-review scale was more uneven than across the five funding bins. Furthermore, the peer-review scores and funding bins were positively and strongly correlated (R = 0.86). Thus, it was decided to use the funding bins as the basis for the comparative analysis.

Table 1 shows the distribution of researchers across funding bins, for each population size. In comparing performance across funding bins, and because the distribution of researchers is somewhat uniform across bins (except the unfunded group, which includes more researchers), this approach enables testing the reliability of the Scopus AUIDs with groups of roughly 20, 100, 200, 500, 1,000 and 1,500 researchers.

The indicators included in this study were the number of papers and the average of relative citations (ARC is similar to MNCS). For both indicators, fractional counting was used. Similar results were obtained using full counting.

Table 1. Distribution of researchers across funding bins and population sizes

Funding bin/ population size	*100*	*500*	*1,000*	*2,500*	*5,000*	*7,500*
0 - Not funded	32 (22, 40)	157 (137, 177)	312 (284, 345)	783 (742, 820)	1,566 (1,519, 1,612)	2,353 (2,312, 2,388)
1 - Lowest funding tier	16 (9, 24)	81 (66, 97)	162 (142, 185)	403 (369, 436)	807 (771, 840)	1,209 (1,179, 1,240)
2 - 2nd funding tier	15 (9, 22)	76 (61, 91)	151 (130, 173)	381 (353, 411)	762 (725, 797)	1,142 (1,111, 1,174)
3 - 3rd funding tier	17 (11, 25)	85 (68, 105)	173 (150, 196)	432 (397, 463)	863 (830, 905)	1,295 (1,263, 1,331)
4 - Highest funding tier	20 (12, 28)	100 (83, 116)	201 (176, 223)	501 (469, 534)	1,001 (962, 1,040)	1,501 (1,466, 1,536)
Avg. across bins	*20*	*100*	*200*	*500*	*1,000*	*1,500*

Results

Recall and precision

Recall and precision are key parameters used in information retrieval to assess the quality of a data set: Were all the right items captured? (This is the recall.) Were all the captured items right? (This is precision.) If the AUID captured 86% of the correct papers—defined here as the manually cleaned portfolio, the ground truth—then the AUID has a recall of 86%. If 5% of papers captured were erroneous, or "false positives," then the AUID has a precision of 95%: 95% of the papers it identified were right.

Across the 10,000 researcher portfolios available to perform this study, the quality of the AUID was very high on average, with a recall of 98% and a precision of 96.9%. However, as sub-population sizes grow smaller and smaller (from right to left in Figure 1), the effects of outliers take on a larger and larger weight within the whole, and thus the smallest sub-populations tended to have the greatest variance in terms of recall and precision (i.e., the widest range [represented by vertical bars in Figure 1] containing 95% of the recall and precision scores across the sampled populations within each population size [i.e., N=500]).[1]

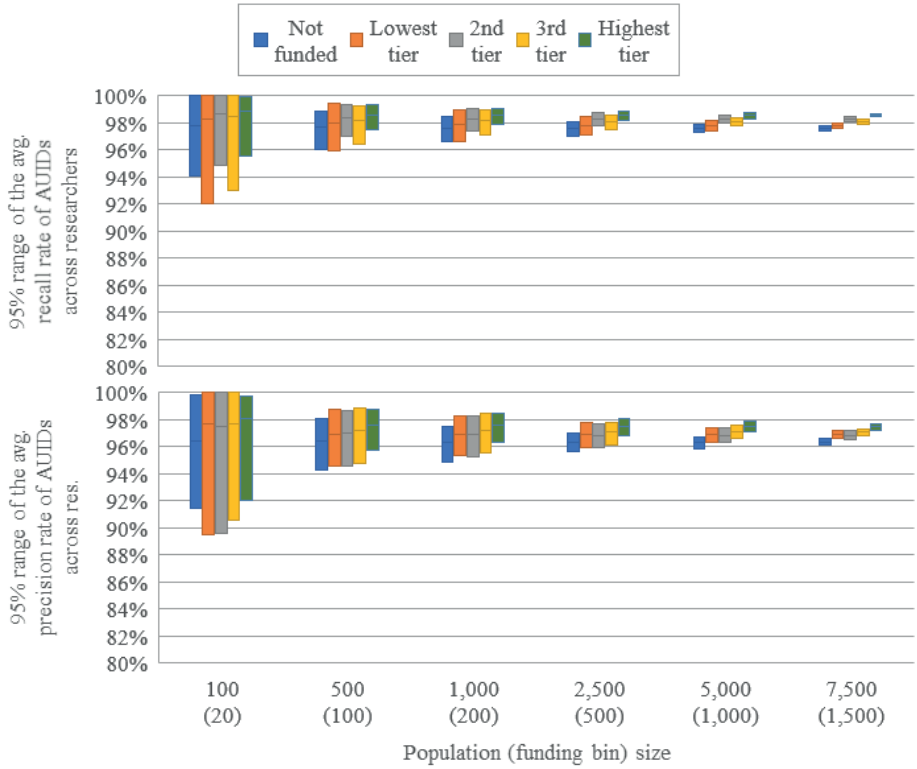


Figure 1. Recall (top) and precision (bottom) of the main Scopus AUID of applicants by population size and funding bin, competition years 2009 to 2013.[2]

From Figure 1, it is obvious that a very high recall and precision (i.e., each above 95%) is achieved with as few as 1,000 researchers, or roughly 200 researchers per funding bin. However, with 500 researchers in total, or 100 per bin, the 95% range of the precision rate widens to include values below 95% and, with 100 researchers, or 20 per bin, performance degrades more noticeably, with the lower limit of precision hovering around 90%. It is thus likely that the use of automated portfolios in comparing performance across funding bins will lead, in some instances with the smaller population sizes, to erroneous conclusions or to an over- or under-estimation of differences between adjacent funding bins.

[1] Throughout the paper, "lower limit" and "upper limit" will be used to denote the boundaries within which 95% of the scores across the 500 samples of each population size are contained. Our intention was to explore most of the variation while excluding the largest outliers that seldom occur within each population size.
[2] Each bar indicates the range containing 95% of the samples' scores.

Guide to reading and interpreting Tables 3 and 4

This section on paper counts and ARC scores explores the impacts of using automated portfolios at different scales, which is to say, with populations of different sizes. This is achieved by comparing the difference in the performance of adjacent funding bins (i.e., comparing applicants who were not funded to those who received the smallest grants, comparing those who received the smallest grants to the next-highest funding bin, and so on) computed using the AUID to the difference computed using manually cleaned data for each of the created populations. The population size where the difference (between the AUID and manually cleaned data) in differences (between funding bins) will converge to 0 will identify the minimal threshold required to achieve reliable results using the AUIDs.

The quantitative assessment of this convergence is presented in Table 2 (paper counts) and Table 3 (ARC). Each data column in these tables presents the median score across the 500 samples, with the lower and upper limits provided in brackets below (noting again that these limits reflect 95% of the samples). The difference in performance between the lower bin and the higher bin is computed based on cleaned portfolios ("Absolute change" column in Table 2 and Table 3), and also expressed as a percent change[3] to better convey the magnitude of the observed difference ("Percent change" column in Table 2 and Table 3).

To assess the fidelity of results one would obtain using the AUID, any bias introduced is presented as both an absolute magnitude and as a percent. The absolute bias is computed as:

$$Absolute\ bias = \left(LBS_{cle} + (HBS_{aut} - LBS_{aut})\right) - HBS_{cle}$$

The percent bias introduced by using automated portfolios is computed as:

$$Percent\ bias = \left(\left(LBS_{cle} + (HBS_{aut} - LBS_{aut})\right)/HBS_{cle}\right) - 1$$

Where,

LBS_{cle} = Lower bin score based on cleaned portfolios,
LBS_{aut} = Lower bin score based on automated portfolios,
HBS_{cle} = Higher bin score based on cleaned portfolios, and
HBS_{aut} = Higher bin score based on automated portfolios.

The biases thus indicate the deviation from the higher bin score based on cleaned portfolios (i.e., the ground truth) that would be obtained using automated portfolios. Specifically, if the cleaned portfolio and the AUID produced the same value for the lower bin, then the bias measures indicate how different the higher bin score would be using the AUID, compared to the higher bin score based on clean data.

For the automated portfolios to provide reliable results, the absolute bias should not be of a magnitude that would be interpreted as a meaningful difference in performance. Additionally, the median percent bias should be close to 0 with a tight 95% interval around that value. The larger the 95% bias interval, the higher the risk of misinterpreting the magnitude of differences between two funding bins at a given population size.

A last column in Table 2 and Table 3 ("Share of population with same effect sign") reports the percentage of populations (i.e., % of the 500 samples at a given population size) for which the difference based on cleaned and automated portfolios shares the same effect sign. When the cleaned and automated portfolios produce effects of a different sign, these would produce inconsistent conclusions in a program evaluation, and thus it is especially important to test for such instances.[4] Given the potential implications (good or bad) of a conclusion opposite to reality for those being assessed in an evaluation context, the confidence level to be achieved

[3] i.e., (higher bin score - lower bin score) / lower bin score

[4] Changes near 0 (i.e., absolute differences between -0.025 and +0.025) were not considered to be of different signs. In such cases, the interpretation for an evaluation would be that the program had negligible effect, and therefore such a small magnitude on either side of 0 was considered to be consistent.

was set very high for this quality dimension: at 100% rather than 95% as for the size effect range.[5] We consider here that a small bias in size effect is acceptable if the conclusion remains in the same direction, whereas a change in effect sign is not. However, while we consider such a rigid threshold to be warranted in an evaluation context—where the stakes are quite high—in other contexts (perhaps oriented more towards exploring the landscape rather than evaluating performance) it is worth considering how rigid a threshold must be respected.

Funding effects measured

Now that the table structure and interpretation have been clarified, let us turn to examining the results in Table 2 and Table 3. Based on cleaned publication portfolios, meaningful increases in paper counts are observed across all funding bin comparisons (from lower to higher funding bins). Across samples, the median increase for papers ranges roughly from 20% to 55% (0.6 to 2.4 in absolute terms; Table 2). For ARC, meaningful increases were only observed between bin 1 and bin 2 as well as between bin 3 and bin 4 (roughly 15% increases around 0.20 in absolute terms; Table 3).

Within most of these comparisons, there was quite a lot of diversity in the magnitude of observed increases within each population size based on the cleaned portfolios. This variation is clearly visible from the range containing 95% of the percent change scores observed across populations of a given size (Table 2 and Table 3). The variation within the cleaned portfolios does not provide any insights about data quality;[6] they are simply performance readings of multiple populations.

For each population created, its performance was computed based on the manually cleaned data and based on the AUID. For instance, for a population of 500 researchers, the performance measurement for population #397 based on clean data was compared to the performance measurement for population #397 based on AUID. Data quality is assessed by the range of biases introduced across these populations. These biases are measured by comparing the performance measurement for, for example, population #397 based on clean data relative to the performance measurement for population #397 based on AUID. These biases will naturally increase with smaller populations due to data quality outliers. The idea is to determine at what population sizes these data quality outliers no longer impact the fidelity of measured differences in scientific performance.

Biases introduced using Scopus AUIDs (i.e., automated portfolios)

Despite this diversity in the magnitude of observed changes in paper counts and ARC, the median percent bias introduced by using automated portfolios converges to near 0% for all funding bin comparisons and population sizes (see Table 2 for paper counts and Table 3 for the ARC). Additionally, while the lower and upper limits of the percent bias show a wide range of results for the populations of 100 researchers—stressing once again that this means only 20 researchers per funding bin—that range reduces rapidly when moving to the larger and larger population sizes. With a population of 5,000 researchers—or 1,000 per bin—the range between the lower and upper limits of bias has already narrowed to no more than ±2%.

[5] i.e., the absolute/percent bias range including 95% of the population scores must be close to 0.

[6] If we were interested in the reliability of random sampling to estimate the parameters of the overall population of 10,000 from which the samples were drawn, the variation across samples based on the cleaned portfolios would inform us on the magnitude of the sampling error at different scales. The analysis of such sampling error is the subject of an ongoing study at Science-Metrix, but it is not explored in this paper

Table 2. Differences in number of papers per researcher moving from lower to higher funding bins based on cleaned researcher portfolios, and bias introduced using Scopus AUIDs

Comparison	*Population size*	*Avg. no. of res. per bin*	*Cleaned portfolios*		*Bias introduced using Scopus AUIDs*		
			Absolute change	*Percent change*	*Absolute bias*	*Percent bias*	*% of pop. w/ same effect sign*
From funding bin 0 (not funded) to 1 (lowest funding tier)	100	20	0.8 (-0.5, 2.2)	37% (-21%, 118%)	0.0 (-0.4, 0.7)	-1% (-15%, 26%)	97%
	500	100	0.9 (0.2, 1.5)	39% (10%, 75%)	0.0 (-0.2, 0.2)	0% (-6%, 8%)	100%
	1,000	200	0.9 (0.4, 1.2)	39% (16%, 60%)	0.0 (-0.1, 0.2)	0% (-5%, 5%)	100%
	2,500	500	0.9 (0.6, 1.1)	39% (26%, 52%)	0.0 (-0.1, 0.1)	0% (-3%, 3%)	100%
	5,000	1,000	0.9 (0.7, 1.0)	39% (31%, 46%)	0.0 (-0.1, 0.0)	0% (-2%, 1%)	100%
	7,500	1,500	0.8 (0.8, 0.9)	39% (34%, 43%)	0.0 (0.0, 0.0)	0% (-1%, 1%)	100%
From funding bin 1 (lowest funding tier) to 2 (2nd funding tier)	100	20	0.7 (-1.4, 2.9)	24% (-32%, 138%)	0.0 (-0.7, 0.7)	1% (-19%, 20%)	94%
	500	100	0.7 (-0.2, 1.5)	21% (-6%, 56%)	0.0 (-0.3, 0.2)	0% (-7%, 7%)	98%
	1,000	200	0.6 (0.1, 1.2)	20% (2%, 44%)	0.0 (-0.2, 0.2)	1% (-4%, 4%)	100%
	2,500	500	0.6 (0.3, 1.0)	21% (10%, 34%)	0.0 (-0.1, 0.1)	0% (-2%, 3%)	100%
	5,000	1,000	0.6 (0.4, 0.8)	21% (14%, 27%)	0.0 (0.0, 0.1)	0% (-1%, 2%)	100%
	7,500	1,500	0.6 (0.5, 0.7)	21% (17%, 25%)	0.0 (0.0, 0.0)	0% (-1%, 1%)	100%
From funding bin 2 (2nd funding tier) to 3 (3rd funding tier)	100	20	0.6 (-1.7, 3.2)	18% (-32%, 114%)	0.0 (-0.6, 0.6)	0% (-14%, 14%)	96%
	500	100	0.8 (-0.3, 1.9)	22% (-6%, 57%)	0.0 (-0.2, 0.2)	0% (-5%, 5%)	100%
	1,000	200	0.8 (0.1, 1.6)	23% (2%, 48%)	0.0 (-0.2, 0.1)	0% (-3%, 3%)	100%
	2,500	500	0.8 (0.4, 1.2)	23% (10%, 34%)	0.0 (-0.1, 0.1)	0% (-2%, 2%)	100%
	5,000	1,000	0.8 (0.6, 1.1)	23% (16%, 30%)	0.0 (-0.1, 0.0)	0% (-1%, 1%)	100%
	7,500	1,500	0.8 (0.7, 1.0)	23% (19%, 27%)	0.0 (0.0, 0.0)	0% (-1%, 0%)	100%

Note: Table continues on next page.

Table 2 (Cont.). Differences in number of papers per researcher moving from lower to higher funding bins based on cleaned researcher portfolios, and bias introduced using Scopus AUIDs

Comparison	*Population size*	*Avg. no. of res. per bin*	*Cleaned portfolios*		*Bias introduced using Scopus AUIDs*		
			Absolute change	*Percent change*	*Absolute bias*	*Percent bias*	*% of pop. w/ same effect sign*
From funding bin 3 (3rd funding tier) to 4 (highest funding tier)	100	20	2.3 (-0.7, 5.7)	55% (-10%, 162%)	0.0 (-0.7, 0.7)	0% (-10%, 11%)	99%
	500	100	2.4 (1.1, 3.9)	54% (23%, 98%)	0.0 (-0.2, 0.2)	0% (-3%, 3%)	100%
	1,000	200	2.4 (1.5, 3.3)	54% (31%, 79%)	0.0 (-0.1, 0.2)	0% (-2%, 2%)	100%
	2,500	500	2.4 (1.9, 3.0)	54% (41%, 70%)	0.0 (-0.1, 0.1)	0% (-1%, 1%)	100%
	5,000	1,000	2.4 (2.1, 2.7)	54% (46%, 63%)	0.0 (0.0, 0.1)	0% (-1%, 1%)	100%
	7,500	1,500	2.4 (2.2, 2.6)	54% (49%, 59%)	0.0 (0.0, 0.0)	0% (0%, 1%)	100%

What would this degree of fidelity look like in practice? Suppose that in comparing bin 0 (not funded) to bin 1 (lowest funding tier), there was a change of 0.8 in number of papers, with bin 0 having a score of 2.2 and bin 1 having a score of 3.0. For the ARC, suppose that no change was measured. The lower bin score was 1.00 and so was the higher bin score.

With a population size of roughly 20 researchers per bin, the range of bias for number of papers was -0.4 to 0.7 in absolute terms, -15% to 26% as a percentage. Applying these numbers to our hypothetical case, using the AUID could lead to a notable overestimation of the effect, measuring a change from 2.2 to 3.8 instead of 2.2 to 3.0. This magnitude of error is highly problematic, as any difference of about 0.5 papers or more per researcher (in fractional counting) would be considered meaningful on the whole population. We can also see that the measured result with the AUID was not always of the same sign as the true figure.

For the ARC, the range of bias observed using the AUID was -0.16 to 0.18 in absolute terms, or -14% to 18% as a percentage. That means that at this scale, using the AUID could lead (using the maximum bias) to an erroneous measurement of change, from 1.00 for bin 0 to 1.18 to bin 1, for example. The magnitude of this change is above the typical threshold for judging that a difference in ARC is meaningful; depending on the volume of publications involved in the measurement, a change between 0.05 to 0.1 is usually considered to be small but approaching a threshold of practical importance. Once again, we see here that the measured result with the AUID was not always of the same sign as the true figure.

From this hypothetical analysis, we can conclude that the use of the AUID on groups of about 20 researchers is not viable. At this population size, for both paper counts and ARC, the differences are notable between the AUID results and the ground truth. The size of the effect can be importantly erroneous, and in 3% of cases the effect sign (increase or decrease) based on the AUID would be opposite to the one based on manually cleaned data.

Table 3. Differences in the average of relative citations moving from lower to higher funding bins based on cleaned researcher portfolios, and bias introduced using Scopus AUIDs

Comparison	*Population size*	*Avg. no. of res. per bin*	*Cleaned portfolios*		*Bias introduced using Scopus AUIDs*		
			Absolute change	*Percent change*	*Absolute bias*	*Percent bias*	*% of pop. w/ same effect sign*
From funding bin 0 (not funded) to 1 (lowest funding tier)	100	20	0.00 (-0.69, 0.80)	0% (-45%, 82%)	0.00 (-0.16, 0.18)	0% (-14%, 18%)	97%
	500	100	0.02 (-0.30, 0.37)	2% (-24%, 36%)	0.00 (-0.13, 0.17)	0% (-11%, 15%)	98%
	1,000	200	0.04 (-0.22, 0.30)	3% (-17%, 28%)	0.00 (-0.07, 0.10)	0% (-5%, 9%)	99%
	2,500	500	0.03 (-0.10, 0.15)	3% (-8%, 13%)	0.00 (-0.04, 0.05)	0% (-3%, 4%)	100%
	5,000	1,000	0.03 (-0.05, 0.10)	3% (-4%, 9%)	0.00 (-0.02, 0.03)	0% (-2%, 2%)	100%
	7,500	1,500	0.03 (0.00, 0.08)	3% (0%, 7%)	0.01 (-0.01, 0.02)	0% (-1%, 1%)	100%
From funding bin 1 (lowest funding tier) to 2 (2nd funding tier)	100	20	0.17 (-0.63, 1.02)	15% (-38%, 117%)	-0.01 (-0.18, 0.12)	-1% (-14%, 10%)	98%
	500	100	0.19 (-0.20, 0.56)	16% (-15%, 54%)	0.00 (-0.17, 0.06)	0% (-13%, 5%)	100%
	1,000	200	0.16 (-0.09, 0.46)	14% (-7%, 44%)	-0.01 (-0.10, 0.04)	0% (-8%, 3%)	100%
	2,500	500	0.18 (0.02, 0.33)	15% (2%, 29%)	0.00 (-0.06, 0.02)	0% (-4%, 1%)	100%
	5,000	1,000	0.17 (0.09, 0.25)	15% (7%, 22%)	-0.01 (-0.03, 0.01)	-1% (-2%, 1%)	100%
	7,500	1,500	0.17 (0.13, 0.22)	15% (10%, 19%)	-0.01 (-0.02, 0.00)	-1% (-2%, 0%)	100%
From funding bin 2 (2nd funding tier) to 3 (3rd funding tier)	100	20	0.04 (-0.70, 0.87)	3% (-39%, 85%)	0.00 (-0.11, 0.18)	0% (-9%, 13%)	99%
	500	100	0.05 (-0.36, 0.39)	4% (-22%, 33%)	0.00 (-0.06, 0.08)	0% (-4%, 5%)	100%
	1,000	200	0.05 (-0.24, 0.27)	4% (-15%, 22%)	0.00 (-0.04, 0.05)	0% (-3%, 3%)	100%
	2,500	500	0.04 (-0.11, 0.17)	3% (-7%, 14%)	0.00 (-0.02, 0.03)	0% (-2%, 2%)	100%
	5,000	1,000	0.04 (-0.03, 0.12)	3% (-2%, 9%)	0.00 (-0.01, 0.02)	0% (-1%, 1%)	100%
	7,500	1,500	0.04 (-0.01, 0.08)	3% (-1%, 6%)	0.00 (-0.01, 0.01)	0% (0%, 1%)	100%

Note: Table continues on next page.

Table 3 (Cont.). Differences in the average of relative citations moving from lower to higher funding bins based on cleaned researcher portfolios, and bias introduced using Scopus AUIDs

Comparison	*Population size*	*Avg. no. of res. per bin*	*Cleaned portfolios*		*Bias introduced using Scopus AUIDs*		
			Absolute change	*Percent change*	*Absolute bias*	*Percent bias*	*% of pop. w/ same effect sign*
From funding bin 3 (3rd funding tier) to 4 (highest funding tier)	100	20	0.19 (-0.72, 1.03)	15% (-36%, 90%)	0.00 (-0.20, 0.10)	0% (-11%, 6%)	98%
	500	100	0.18 (-0.16, 0.78)	13% (-10%, 56%)	0.00 (-0.07, 0.04)	0% (-4%, 3%)	100%
	1,000	200	0.20 (-0.07, 0.56)	14% (-5%, 43%)	0.00 (-0.05, 0.03)	0% (-3%, 2%)	100%
	2,500	500	0.20 (0.05, 0.39)	14% (4%, 29%)	0.00 (-0.03, 0.02)	0% (-2%, 1%)	100%
	5,000	1,000	0.20 (0.11, 0.30)	15% (8%, 22%)	-0.01 (-0.02, 0.01)	0% (-1%, 0%)	100%
	7,500	1,500	0.21 (0.14, 0.26)	15% (10%, 19%)	-0.01 (-0.01, 0.00)	0% (-1%, 0%)	100%

With larger and larger populations, though, the magnitudes of the biases decrease. Had we used the AUID on a population of roughly 500 researchers per bin, the range of bias for paper counts would have topped out at around 3%. In practical terms, this means that with the AUID we could have measured a change from 2.2 to 3.1 instead of the real change of 2.2 to 3.0. Given that a threshold for meaningful difference is around 0.5 on this indicator, the error here is of tolerable magnitude. Furthermore, in 100% of cases, the effect would at least be of the correct sign, even if the magnitude is slightly off.

Turning to the ARC for a population of 500 researchers per bin, the upper limit of bias observed was 4%, meaning that we would have measured a change from 1.00 to 1.04, just about at the lower limit of a small, yet meaningful, difference in performance, which is 0.05. Here also we see that in 100% of cases, the effect would at least be of the correct sign, though the magnitude of error is right on the lower margin where the measured value would possibly be interpreted as meaningfully different from the real value.

From these analyses, it seems like the AUID can be used to compare groups of about 500 researchers. However, it would be wise to exercise caution in interpreting the findings here by slightly raising one's normal thresholds for judging a change to be meaningful. Also, the fact that the results for paper counts were very reliable and the results for ARC called for some caution highlights an interesting point: this analysis of data quality for use case needs to consider which indicators will be computed on this basis, how sensitive those indicators are to underlying variation in the data, and where their respective thresholds are for judging an effect to be meaningful (which is also context dependent).

What if we look at an even larger population—say, 1,000 researchers per bin? Here the percent bias for papers would not exceed 2%, so based on the AUID one would measure a change from 2.2 to 3.07, very close to the one measured using the manually cleaned profiles (i.e., 3.01). Similarly, the percent bias for the ARC would not exceed 2%, so based on the AUID one would measure a change from 1.00 to 1.02, whereas the clean data showed no change whatsoever—that is, the AUID would show a negligible change rather than no change. For both papers and ARC, the measured result would be of the correct sign in 100% of cases.

Discussion

The study indicates that the use of AUIDs to automatically disambiguate the papers of researchers in Scopus is adequate for comparing groups of approximately 500 researchers or more (i.e., population size of 2,500 researchers in this study with an average production of at least 3 papers per researcher based on fractional counting).[7] For the tested populations that were smaller than that (i.e., roughly 200 researchers or fewer per bin with an average production of at least 3 papers per researcher), the Scopus AUIDs do not always provide reliable conclusions. Specifically, measurements based on the AUID may suggest an effect size meaningfully different than the real effects observed based on manual cleaning. With groups of 200 or fewer researchers, using the AUID may even yield, on few occasions, a result of a different sign.

In practical terms, such errors would lead a program evaluation to conclude that there was no effect when in fact there was one, that there was an effect when in fact there was not, or even that there was a negative effect when in fact the effect was positive (or vice versa). Given that the quality of the AUID is being tested here for use cases such as this one, the results show that the AUID can pass this test only when using populations of 500 or more researchers per group (with an average production of at least 3 papers per researcher).

Of course, while the group size of 500 researchers represents something of a turning point in terms of reliability of results, using a group of an even larger size would provide a buffer. Such a buffer would be wisely implemented, given that this study has not been replicated, that it focuses on North American researchers, that it focuses on the natural and applied sciences, and that it focuses on a single funder. Further research may demonstrate that a different threshold of population size is necessary to obtain satisfactory quality, especially in other contexts.

Nonetheless, given the experience of the study team in working with bibliometric data, we are confident putting forward the hypothesis that the present findings are most likely to be generalizable to contexts that are similar along various dimensions. It is likely safe to assume that similar thresholds could be applied to North American researchers in the health sciences as well as to European researchers in either the natural and applied sciences or the health sciences. Issues related to the disambiguation of author names in bibliographic databases (e.g., homonyms, typos) are typically not more pronounced for European than North American names. Additionally, the volume of production per researcher in the health sciences should be large enough to derive similar conclusions regarding the scale (in terms of population/group size) at which the AUIDs become sufficiently reliable for application in research evaluation.

By contrast, there are additional challenges to the disambiguation of Asian names: homonyms are much more problematic with Asian names, which often leads to automated author identifiers of lower quality (especially in terms of precision). Similarly, there are unique challenges to measuring activity in the social sciences and in the humanities: the volume of production in those areas is typically much lower than it is in the natural, applied and health sciences in databases such as Scopus and the Web of Science. As a result, further research would be needed to identify a safe population size for using the AUID for evaluations in such contexts.

Accounting for the above conclusions, we put forward three scenarios to guide the proper use of Scopus AUIDs in a research evaluation context dealing with the comparative analysis of various groups of researchers. Recall that by using Scopus AUIDs, our main goal is to offer cost-effective and reliable solutions to research funders when dealing with very large populations of researchers. The scenarios for using the Scopus AUIDs are as follows:[8]

---

[7] The volume of papers per researcher in a group also has an influence on the threshold where the AUID becomes sufficiently reliable.

[8] The thresholds established in the following scenarios in terms of number of researchers assume an average production of at least 3 papers per researcher based on fractional counting (8 papers based on full counting).

1. **Small populations (fewer than 500 researchers per group):** Proceed with the manual cleaning of the researchers' publication data.
2. **Mid-range populations (around 500 researchers per group):** If the client can afford to manually clean the publication data of researchers, then proceed with this approach to accommodate a buffer in this "transition" zone (in terms of group sizes). If not, then the conclusions using Scopus AUIDs should be reliable provided that the studied population shares similar characteristics to this study's population: North American or European geography; natural, applied or health science thematic areas. That said, the size effect could be biased by about +/- 2%–4% (depending on indicator and compared groups). Care should be taken not to overinterpret the findings obtained; it would be prudent to set a higher threshold for differences to be considered meaningful.
3. **Large populations (around or more than 1,000 researchers per group):** The use of Scopus AUIDs is warranted to decrease study costs since they provide highly reliable conclusions both in terms of direction and size of an effect. Here again, the studied population should share similar characteristics to this study's population.

When a very large study population includes relevant sub-groups of varying sizes (e.g., less than 500, around 500, 1,000 or more per group), the study design could include a mixed-method approach. To optimize the resulting value of the evaluation while minimizing its costs, it is recommended to make use of the Scopus AUIDs where the groups being compared are large enough, and to manually clean the data otherwise. The choice of methods would be guided according to the thresholds established in point 1 to 3 above.

Acknowledgments

Although Science-Metrix was recently acquired by Elsevier, which produces Scopus, the results of this study were not altered in any way to portray the AUID in a more favourable light, nor were any shortcomings identified that have not been disclosed here. The study authors also wish to acknowledge the exceptional work of all the bibliometricians who contributed to the manual construction of well curated publication portfolios for the 10,000 researchers used for this study.

References

Archambault, É., Beauchesne, O. H., & Caruso, J. (2011). Towards a multilingual, comprehensive and open scientific journal ontology. In B. Noyons, P. Ngulube, & J. Leta (Eds.), *Proceedings of the 13th International Conference of the International Society for Scientometrics and Informetrics (ISSI)* (pp. 66–77). Durban, South Africa.

D'Angelo, C. A., Giuffrida, C., & Abramo, G. (2011). A heuristic approach to author name disambiguation in bibliometrics databases for large-scale research assessments. *Journal of the American Society for Information Science and Technology*, *62*(2), 257–269.

Gurney, T., Horlings, E., & van den Besselaar, P. (2012). Author disambiguation using multi-aspect similarity indicators. *Scientometrics*, *91*(2), 435–449.

Picard-Aitken, M., Foster, T., Labrosse, I., Caruso, J., Campbell, D., & Archambault, É. (2010). *Tenth-year evaluation of the Canada Research Chairs Program*. Prepared for the Social Sciences and Humanities Research Council of Canada.

Science-Metrix. (2018). *Evaluation of the Marine Microbiology Initiative*. Retrieved from https://www.moore.org/docs/default-source/mmi-documents/marine-microbiology-initiative-evaluation9f820661a10f68a58452ff00002785c8.pdf?sfvrsn=c58d6e0c_0

Strotmann, A., & Zhao, D. (2012). Author name disambiguation: What difference does it make in author-based citation analysis? *Journal of the American Society for Information Science and Technology*, *63*(9), 1820–1833. https://doi.org/10.1002/asi.22695

Tang, L., & Walsh, J. P. (2010). Bibliometric fingerprints: name disambiguation based on approximate structure equivalence of cognitive maps. *Scientometrics*, *84*(3), 763–784. https://doi.org/10.1007/s11192-010-0196-6

Intermediacy of publications

Lovro Šubelj[1], Ludo Waltman[2], Vincent Traag[2] and Nees Jan van Eck[2]
[1]Faculty of Computer and Information Science, University of Ljubljana, Ljubljana, Slovenia
[2]Centre for Science and Technology Studies, Leiden University, Leiden, Netherlands

Abstract

Citation networks of scientific publications offer fundamental insights into the structure and development of scientific knowledge. We propose a new measure, called intermediacy, for tracing the historical development of scientific knowledge. Given two publications, an older and a more recent one, intermediacy identifies publications that seem to play a major role in the historical development from the older to the more recent publication. The identified publications are important in connecting the older and the more recent publication in the citation network. After providing a formal definition of intermediacy, we study its mathematical properties. We then present two empirical case studies, one tracing historical developments at the interface between the community detection and the scientometric literature and one examining the development of the literature on peer review. We show both mathematically and empirically how intermediacy differs from main path analysis, which is the most popular approach for tracing historical developments in citation networks. Main path analysis tends to favor longer paths over shorter ones, whereas intermediacy has the opposite tendency. Compared to main path analysis, we conclude that intermediacy offers a more principled approach for tracing the historical development of scientific knowledge.

Introduction

Citation networks provide invaluable information for tracing historical developments in science. The idea of tracing scientific developments based on citation data goes back to Eugene Garfield, the founder of the Science Citation Index. In a report published more than 50 years ago, Garfield and his co-workers concluded that citation analysis is “a valid and valuable means of creating accurate historical descriptions of scientific fields” (Garfield, Sher, & Torpie, 1964). Garfield also developed a software tool called HistCite that visualizes citation networks of scientific publications. This tool supports users in tracing historical developments in science, a process sometimes referred to as algorithmic historiography by Garfield (Garfield, Pudovkin, & Istomin, 2003b; Garfield, Pudovkin, & Istomin, 2003a; Garfield, 2004). More recently, a software tool called CitNetExplorer (van Eck & Waltman, 2014) was developed that has similar functionality but offers more flexibility in analyzing large-scale citation networks. Other software tools, most notably CiteSpace (Chen, 2006) and CRExplorer (Marx, Bornmann, Barth, & Leydesdorff, 2014; Thor, Marx, Leydesdorff, & Bornmannd, 2016), provide alternative approaches for tracing scientific developments based on citation data.

Main path analysis, originally proposed by Hummon and Doreian (Hummon & Doreian, 1989), is a widely used technique for tracing historical developments in science. Given a citation network, main path analysis identifies one or more paths in the network that are considered to represent the most important scientific developments. Many variants and extensions of main path analysis have been proposed (Batagelj, 2003; Lucio-Arias & Leydesdorff, 2008; Liu & Lu, 2012; Batagelj, Doreian, Ferligoj, & Kejžar, 2014; Yeo, Kim, Lee, & Kang, 2014; Liu & Kuan, 2016; Tu & Hsu, 2016), not only for citation networks of scientific publications but also for patent citation networks (Verspagen, 2007; Park & Magee, 2017; Gwak & Sohn, 2018; Kim & Shin, 2018; Kuan, Huang, & Chen, 2018).

In this paper, we introduce a new approach for tracing historical developments in science based on citation networks. We propose a measure called intermediacy. Given two publications dealing with a specific research topic, an older publication and a more recent one, intermediacy can be used to identify publications that appear to play a major role in the historical development from the older to the more recent publication. These are publications

that, based on citation links, are important in connecting the older and the more recent publication.
Like main path analysis, intermediacy can be used to identify one or more citation paths between two publications. However, as we will make clear, there are fundamental differences between intermediacy and main path analysis. Most significantly, we will show that main path analysis tends to favor longer citation paths over shorter ones, whereas intermediacy has the opposite tendency. For the purpose of tracing historical developments in science, we argue that intermediacy yields better results than main path analysis.

Intermediacy

Consider a directed acyclic graph $G=(V,E)$, where V denotes the set of nodes of G and E denotes the set of edges of G. The edges are directed. We are interested in the connectivity between a source $s \in V$ and a target $t \in V$. Only nodes that are located on a path from source s to target t are of relevance. We refer to such a path as a source-target path. We assume that each node $v \in V$ is located on a source-target path.

Definition 1. Given a source s and a target t, a path from s to t is called a source-target path.

In this paper, our focus is on citation networks of scientific publications. In this context, nodes are publications and edges are citations. We choose edges to be directed from a citing publication to a cited publication. Hence, edges point backward in time. This means that the source is a more recent publication and the target an older one.
Informally, the more important the role of a node $v \in V$ in connecting source s to target t, the higher the intermediacy of v. To formally define intermediacy, we assume that each edge $e \in E$ is active with a certain probability p. We assume that the probability of being active is the same for all edges $e \in E$. Based on the idea of active and inactive edges, we introduce the following definitions.

Definition 2. If all edges on a path are active, the path is called active. Otherwise the path is called inactive. If a node $v \in V$ is located on an active source-target path, the node is called active. Otherwise the node is called inactive.

For two nodes $u, v \in V$, we use X_{uv} to indicate whether there is an active path (or multiple active paths) from node u to node v ($X_{uv}=1$) or not ($X_{uv}=0$). The probability that there is an active path from node u to node v is denoted by $\Pr(X_{uv}=1)$. We use $X_{st}(v)$ to indicate whether there is an active source-target path that goes through node v ($X_{st}(v)=1$) or not ($X_{st}(v)=0$). The probability that there is an active source-target path that goes through node v is denoted by $\Pr(X_{st}(v)=1)=\Pr(X_{sv}=1)\Pr(X_{vt}=1)$. This probability equals the probability that node v is active.
Intermediacy can now be defined as follows.

Definition 3. The intermediacy ϕ_v of a node $v \in V$ is the probability that v is active, that is,

$$\phi_v = Pr(X_{st}(v)=1) = Pr(X_{sv}=1)\, Pr(X_{vt}=1). \tag{1}$$

In the interpretation of intermediacy, we focus on the ranking of nodes relative to each other. We do not consider the absolute values of intermediacy. For instance, suppose the intermediacy of node $v \in V$ is twice as high as the intermediacy of node $u \in V$. We then consider node v to be more important than node u in connecting the source s and the target t. However, we do not consider node v to be twice as important as node u. We now present an

analysis of the mathematical properties of intermediacy. For brevity, we do not include the proofs of the mathematical results, but they can be found in the full paper on https://arxiv.org/abs/1812.08259.

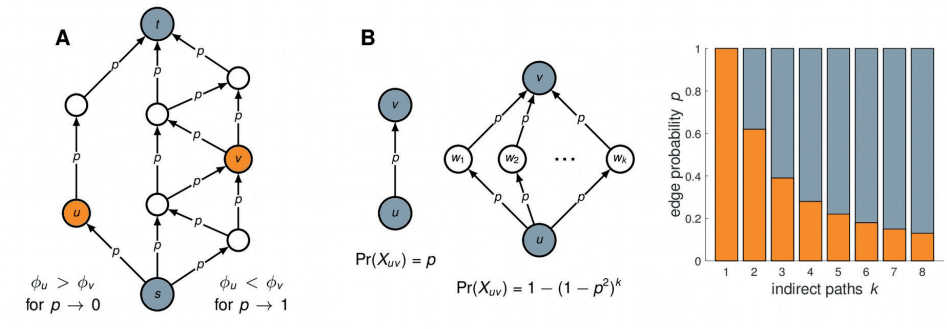


Figure 1: (A) Illustration of the limit behavior of intermediacy. For $p \to 0$, intermediacy favors nodes located on shorter paths and therefore node u has a higher intermediacy than node v. For $p \to 1$, intermediacy favors nodes located on a larger number of edge independent paths and therefore node v has a higher intermediacy than node u. (B) Illustration of the choice of the parameter p. Nodes u and v are connected by a single direct path in the left graph and by k indirect paths of length 2 in the right graph. For different values of k, the bar chart shows the values of p for which the probability that there is an active path from node u to node v is higher (in orange) or lower (in gray) in the left graph than in the right graph.

Limit behavior

To get a better understanding of intermediacy, we study the behavior of intermediacy in two limit cases, namely the case in which the probability p that an edge is active goes to 0 and the case in which the probability p goes to 1. In each of the two cases, the ranking of the nodes in a graph based on intermediacy turns out to have a natural interpretation. The difference between the two cases is illustrated in Fig. 1A.

Let l_v denote the length of the shortest source-target path going through node $v \in V$. The following theorem states that in the limit as the probability p that an edge is active tends to 0, the ranking of nodes based on intermediacy coincides with the ranking based on l_v. Nodes located on shorter source-target paths are more intermediate than nodes located on longer source-target paths.

Theorem 1. In the limit as the probability p tends to 0, $l_u < l_v$ implies $\phi_u > \phi_v$.

The intuition underlying this theorem is as follows. When the probability that an edge is active is close to 0, almost all edges are inactive. Consequently, almost all source-target paths are inactive as well. However, from a relative point of view, longer source-target paths are more likely to be inactive than shorter source-target paths. This means that nodes located on shorter source-target paths are more likely to be active than nodes located on longer source-target paths (even though for all nodes the probability of being active is close to 0). Nodes located on shorter source-target paths therefore have a higher intermediacy than nodes located on longer source-target paths.

We now consider the limit case in which the probability p that an edge is active goes to 1. Let σ_v denote the number of edge independent source-target paths going through node $v \in V$. Theorem 2 states that in the limit as p tends to 1, the ranking of nodes based on intermediacy

coincides with the ranking based on σ_v. The larger the number of edge independent source-target paths going through a node, the higher the intermediacy of the node.

Theorem 2. In the limit as the probability p tends to 1, $\sigma_u > \sigma_v$ implies $\phi_u > \phi_v$.

Intuitively, this theorem can be understood as follows. When the probability that an edge is active is close to 1, almost all edges are active. Consequently, almost all source-target paths are active as well, and so are almost all nodes. A node is inactive only if all source-target paths going through the node are inactive. If there are σ edge independent source-target paths that go through a node, this means that the node can be inactive only if there are at least σ inactive edges. Consider two nodes $u, v \in V$. Suppose that the number of edge independent source-target paths going through node v is larger than the number of edge independent source-target paths going through node u. In order to be inactive, node v then requires more inactive edges than node u. This means that node v is less likely to be inactive than node u (even though for both nodes the probability of being inactive is close to 0). Hence, node v has a higher intermediacy than node u. More generally, nodes located on a larger number of edge independent source-target paths have a higher intermediacy than nodes located on a smaller number of edge independent source-target paths.
Values of p between 0 and 1 interpolate in some way between the two extremes. For lower values of p the path lengths are relatively more important, whereas for higher values of p the number of edge independent paths are more important. The balance between the two extremes is illustrated in Fig. 1B which compares the case of a path of length 1 to the case of k edge independent paths of length 2.

Path addition and contraction

Next, we study two additional properties of intermediacy, the property of path addition and the property of path contraction. We show that both adding paths and contracting paths lead to an increase in intermediacy. Path addition and path contraction are important properties because they reflect the basic intuition underlying the idea of intermediacy.
We start by considering the property of path addition. We define path addition as follows.

Definition 4. Consider a directed acyclic graph $G=(V,E)$ and two nodes $u, v \in V$ such that there does not exist a path from node v to node u. Path addition is the operation in which a new path from node u to node v is added. Let l denote the length of the new path. If $l=1$, an edge (u,v) is added. If $l>1$, nodes $w_1, \ldots, w_{l-1}$ and edges (u, w_1), (w_1, w_2), …, (w_{l-2}, w_{l-1}), (w_{l-1}, v) are added.

This definition includes the condition that there does not exist a path from node v to node u. This ensures that the graph G will remain acyclic after adding a path. The following theorem states that adding a path increases intermediacy.

Theorem 3. Consider a directed acyclic graph $G=(V,E)$, a source $s \in V$, and a target $t \in V$. In addition, consider two nodes $u, v \in V$ such that there does not exist a path from node v to node u. Adding a path from node u to node v increases the intermediacy ϕ_w of any node $w \in V$ located on a path from source s to node u or from node v to target t.

Theorem 3 does not depend on the probability p. Adding a path always increases intermediacy, regardless of the value of p. To illustrate the theorem, consider Fig. 2A and Fig. 2B. The graph in Fig. 2B is identical to the one in Fig. 2A except that a path from node u to

node v has been added. As can be seen, adding this path has increased the intermediacy of nodes located between source s and node u or between node v and target t, including nodes u and v themselves. While the intermediacy of other nodes has not changed, the intermediacy of these nodes has increased from 0.17 to 0.23. This reflects the basic intuition that, after a path from node u to node v has been added, going from source s to target t through nodes u and v has become 'easier' than it was before. This means that nodes located between source s and node u or between node v and target t have become more important in connecting the source and the target. Consequently, the intermediacy of these nodes has increased.

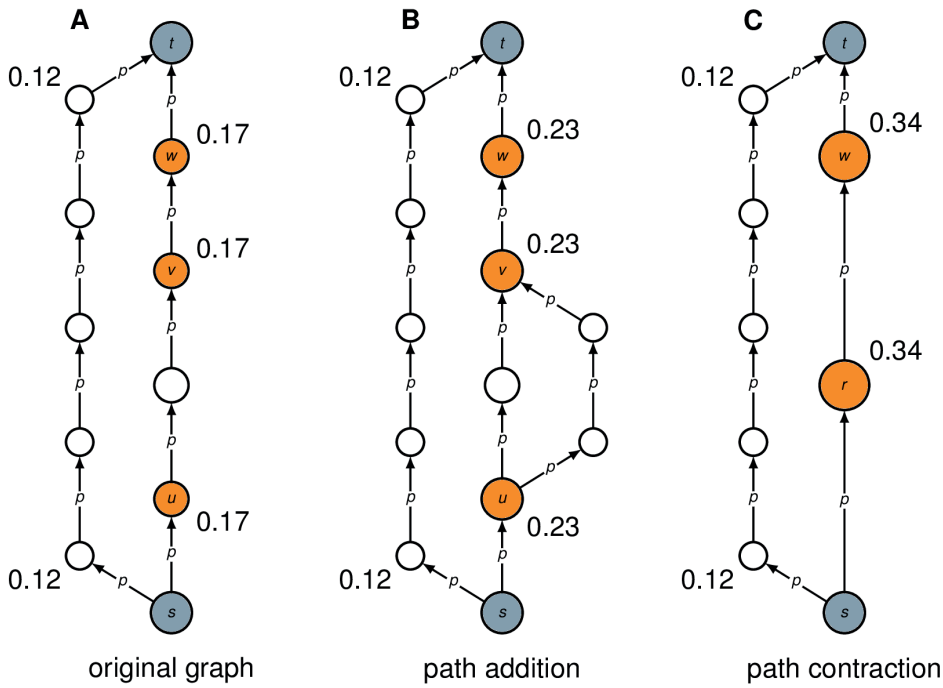


Figure 2: Illustration of the properties of path addition and path contraction. Comparing (B) to (A) shows how path addition increases intermediacy. Comparing (C) to (B) shows how path contraction increases intermediacy. For some nodes in (A), (B), and (C), the intermediacy is reported, calculated using a value of 0.7 **for the probability** p**.**

We now consider the property of path contraction. We use V_{uv} to denote the set of all nodes located on a path from node u to node v, including nodes u and v themselves. Path contraction is then defined as follows.

Definition 5. Consider a directed acyclic graph $G=(V,E)$ and two nodes $u,v \in V$ such that there exists at least one path from node u to node v. Path contraction is the operation in which all nodes in V_{uv} are contracted. This means that the nodes in V_{uv} are replaced by a new node r. Edges pointing from a node $w \notin V_{uv}$ to nodes in V_{uv} are replaced by a single new edge (w,r). Edges pointing from nodes in V_{uv} to a node $w \notin V_{uv}$ are replaced by a single new edge (r,w). Edges between nodes in V_{uv} are removed.

The following theorem states that contracting paths increases intermediacy.

Theorem 4. Consider a directed acyclic graph $G=(V,E)$, a source $s \in V$, and a target $t \in V$. In addition, consider two nodes $u, v \in V$ such that there exists at least one path from node u to node v and such that nodes in V_{uv} do not have neighbors outside V_{uv} except for incoming neighbors of node u and outgoing neighbors of node v. Contracting paths from node u to node v increases the intermediacy ϕ_w of any node $w \in V$ located on a path from source s to node u or from node v to target t.

Like Theorem 3, Theorem 4 does not depend on the probability p. Theorem 4 is illustrated in Fig. 2B and Fig. 2C. The graph in Fig. 2C is identical to the one in Fig. 2B except that paths from node u to node v have been contracted. As a result, there has been an increase in the intermediacy of nodes located between source s and node u or between node v and target t, including nodes u and v themselves (which have been contracted into a new node r). While the intermediacy of other nodes has not changed, the intermediacy of these nodes has increased from 0.23 to 0.34. This reflects the basic intuition that, after paths from node u to node v have been contracted, going from source s to target t through nodes u and v has become 'easier' than it was before. In other words, nodes located on a path from source s to target t going through nodes u and v have become more important in connecting the source and the target, and hence the intermediacy of these nodes has increased.

Alternative approaches

How does intermediacy differ from alternative approaches? We consider two alternative approaches. One is main path analysis (Hummon & Doreian, 1989). This is the most commonly used approach for tracing the historical development of scientific knowledge in citation networks. The other alternative approach is the expected path count approach. Like intermediacy, the expected path count approach distinguishes between active and inactive edges and focuses on active source-target paths. While intermediacy considers the probability that there is at least one active source-target path going through a node, the expected path count approach considers the expected number of active source-target paths that go through a node.

Consider the graph shown in Fig. 3A. To get from source s to target t, one could take either a path going through nodes u and v or the path going through node w. Based on intermediacy, the latter path represents a stronger connection between the source and the target than the former one. This follows from the path contraction property.

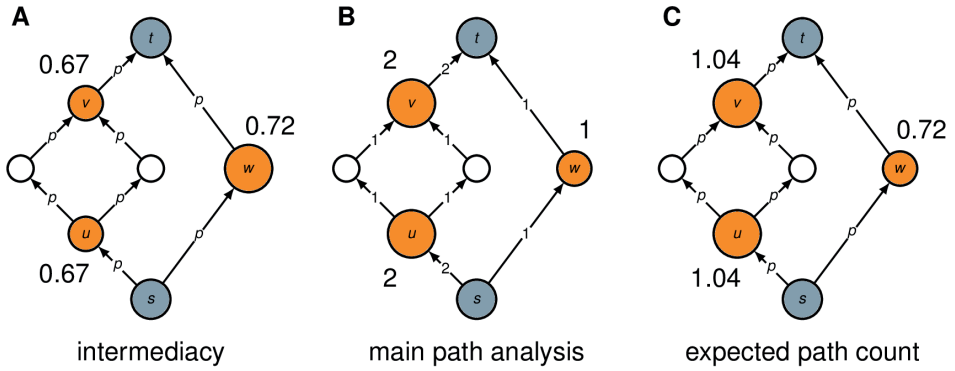


Figure 3: Comparison of intermediacy (A), main path analysis (B), and expected path count (C). For nodes u, v, and w, the intermediacy (A), path count (B), and expected path count (C) are reported, using a value of 0.85 for the probability p in the calculation of intermediacy and expected path count.

Interestingly, main path analysis gives the opposite result, as can be seen in Fig. 3B. For each edge, the figure shows the search path count, which is the number of source-target paths that go through the edge. There are two source-target paths that go through (s,u) and (v,t), while all other edges are included only in a single source-target path. Because the search path counts of (s,u) and (v,t) are higher than the search path counts of (s,w) and (w,t), main path analysis favors paths going through nodes u and v over the path going through node w. This is exactly opposite to the result obtained using intermediacy. Fig. 3B makes clear that main path analysis yields outcomes that violate the path contraction property. Main path analysis tends to favor longer paths over shorter ones. For the purpose of identifying publications that play an important role in connecting an older and a more recent publication, we consider this behavior to be undesirable. There are various variants of main path analysis, which all show the same type of undesirable behavior.

Instead of focusing on the probability of the existence of at least one active source-target path, as is done by intermediacy, one could also focus on the expected number of active source-target paths going through a node. This alternative approach, which we refer to as the expected path count approach, is illustrated in Fig. 3C. As can be seen in the figure, nodes u and v have a higher expected path count than node w. Paths going through nodes u and v may therefore be favored over the path going through node w. Fig. 3C shows that, unlike intermediacy, the expected path count approach does not have the path contraction property. Depending on the probability p, contracting paths may cause expected path counts to decrease rather than increase. Because the expected path count approach does not have the path contraction property, we do not consider this approach to be a suitable alternative to intermediacy.

Empirical analysis

We now present two case studies that serve as empirical illustrations of the use of intermediacy. Case 1 deals with the topic of community detection and its relationship with scientometric research. This case was selected because we are well acquainted with the topic. Case 2 deals with the topic of peer review. This case is of interest because it was recently examined using main path analysis by Batagelj, Ferligoj, & Squazzoni (2017). In both case studies, the intermediacy of publications was calculated using a Monte Carlo algorithm we developed (available at https://github.com/lovre/intermediacy, Šubelj, 2018).

Case 1: Community detection and scientometrics

We analyze how a method for community detection in networks ended up being used in the field of scientometrics to construct classification systems of scientific publications. In particular, we are interested in the development from Newman and Girvan (2004) to Klavans and Boyack (2017). These are our target and source publications. Newman and Girvan (2004) introduced a new measure for community detection in networks, known as modularity, while Klavans and Boyack (2017) compared different ways in which modularity-based approaches can be used to identify communities in citation networks.

Our analysis relies on data from the Scopus database produced by Elsevier. We also considered the Web of Science database produced by Clarivate Analytics. However, many citation links relevant for our analysis are missing in Web of Science. There are also missing citation links in Scopus, but for Scopus the problem is less significant than for Web of Science. We refer to Van Eck and Waltman (2017) for a further discussion of the problem of missing citation links.

In the Scopus database, we found $n=64{,}223$ publications that are located on a citation path between our source and target publications. In total, we identified $m=280{,}033$ citation links

between these publications. This means that on average each publication has $k=2m/n\approx 8.72$ citation links, counting both incoming and outgoing links.
Based on our expert knowledge of the topic under study, we found that the most useful results were obtained by setting the parameter p equal to 0.1. Table 1 lists the ten publications with the highest intermediacy for $p=0.1$. For each publication, the intermediacy is reported for five different values of p. In addition, the table also reports each publications citation count and reference count. Fig. 4A shows the citation network of the ten most intermediate publications for $p=0.1$.

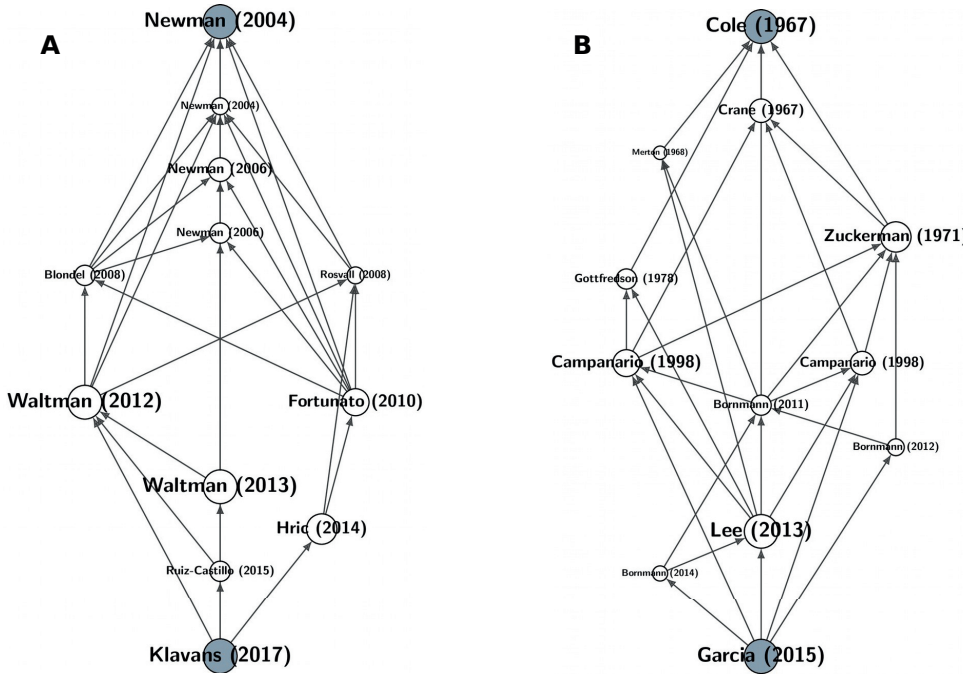


Figure 4: Citation network of the top ten most intermediate publications for $p=0.1$ are shown for (A) case 1 and (B) case 2. (Only the name of the first author is shown.)

Using our expert knowledge to interpret the results presented in Table 1 and Fig. 4A, we are able to trace how a method for community detection ended up in the scientometric literature. The two publications with the highest intermediacy (Waltman & Van Eck, 2012, 2013) played a key role in introducing modularity-based approaches in the scientometric community. Waltman and Van Eck (2012) proposed the use of modularity-based approaches for constructing classification systems of scientific publications, while Waltman and Van Eck (2013) introduced an algorithm for implementing these modularity-based approaches. This algorithm can be seen as an improvement of the so-called Louvain algorithm introduced by Blondel et al. (2008), which is also among the ten most intermediate publications. Most of the other publications in Table 1 and Fig. 4A are classical publications on community detection in general and modularity in particular. The publications by Newman all deal with modularity-based community detection. Rosvall and Bergstrom (2008) proposed an alternative approach to community detection. They applied their approach to a citation network of scientific journals, which explains the connection with the scientometric literature. Fortunato (2010) is a review of the literature on community detection. The intermediacy of this publication is probably strongly influenced by its large number of references. Hric et al. (2014) is a more

recent publication on community detection. This publication focuses on the challenges of evaluating the results produced by community detection methods. This issue is very relevant in a scientometric context, and therefore the publication was cited by our source publication (Klavans & Boyack, 2017). Finally, there is one more scientometric publication in Table 1 and Fig. 4A. This publication (Ruiz-Castillo & Waltman, 2015) is one of the first studies presenting a scientometric application of classification systems of scientific publications constructed using a modularity-based approach. The publication was also cited by our source publication.

Table 1: Top ten most intermediate publications in case 1 for $p=0.1$.

		ϕ	cit.	ref.
t	Newman & Girvan (2004), Finding and evaluating community structure in networks, *Phys. Rev. E* **69**(2), 026113.	0.301	468	0
s	Klavans & Boyack (2017), Which type of citation analysis generates the most accurate taxonomy of scientific and technical knowledge?, *J. Assoc. Inf. Sci. Tec.* **68**(4), 984-998.	0.301	0	24
1	Waltman & Van Eck (2013), A smart local moving algorithm for large-scale modularity-based community detection, *Eur. Phys. J. B* **86**, 471.	0.061	2	27
2	Waltman & Van Eck (2012), A new methodology for constructing a publication-level classification system of science, *J. Assoc. Inf. Sci. Tec.* **63**(12), 2378-2392.	0.060	15	22
3	Hric et al. (2014), Community detection in networks: Structural communities versus ground truth, *Phys. Rev. E* **90**(6), 062805.	0.052	1	29
4	Fortunato (2010), Community detection in graphs, *Phys. Rep.* **486**(3-5), 75-174.	0.037	73	154
5	Newman (2006), Modularity and community structure in networks, *P. Natl. Acad. Sci. USA* **103**(23), 8577-8582.	0.035	221	8
6	Ruiz-Castillo & Waltman (2015), Field-normalized citation impact indicators using algorithmically constructed classification systems of science, *J. Informetr.* **9**(1), 102-117.	0.024	2	24
7	Blondel et al. (2008), Fast unfolding of communities in large networks, *J. Stat. Mech.*, P10008.	0.022	78	21
8	Newman (2006), Finding community structure in networks using the eigenvectors of matrices, *Phys. Rev. E* **74**(3), 036104.	0.021	138	18
9	Newman (2004), Fast algorithm for detecting community structure in networks, *Phys. Rev. E* **69**(6), 066133.	0.020	246	1
10	Rosvall & Bergstrom (2008), Maps of random walks on complex networks reveal community structure, *P. Natl. Acad. Sci. USA* **105**(4), 1118-1123.	0.020	70	10

The citation counts reported in Table 1 show that some publications, especially the more recent ones, have a high intermediacy even though they have been cited only a very limited number of times. This makes clear that a ranking of publications based on intermediacy is quite different from a citation-based ranking of publications. The publications in Table 1 that have a high intermediacy and a small number of citations do have a substantial number of references.

Case 2: Peer review

We now turn to case 2, in which we analyze the literature on peer review. The analysis is based on data from the Web of Science database. We make use of the same data that was also used in a recent paper by Batagelj et al. (2017).

We started with a citation network of 45,965 publications dealing with peer review. This is the citation network that was labeled CiteAcy by Batagelj et al. (2017). We selected Cole and Cole (1967) and Garcia et al. (2015) as our target and source publications. The main path analysis carried out by Batagelj et al. (2017) suggests that these are central publications in the literature on peer review. For the purpose of our analysis, only publications located on a citation path between our source and target publications are of relevance. Other publications

play no role in the analysis. We therefore restricted the analysis to the $n=615$ publications located on a citation path from Garcia et al. (2015) to Cole and Cole (1967). These publications are connected by $m=3420$ citation links, resulting in an average of $k=2m/n\approx 11.12$ citation links per publication.

Table 2 lists the ten publications with the highest intermediacy, where we use a value of 0.1 for the parameter p, like in Table 1. Fig. 4B shows the citation network of the ten most intermediate publications. There are numerous paths in this citation network going from our source publication (Garcia et al., 2015) to our target publication (Cole & Cole, 1967). We regard these paths as the core paths between the source and target publications.

Table 2: Top ten most intermediate publications in case 2 for $p=0.1$.

		ϕ	cit.	ref.
t	Cole & Cole (1967), Scientific output and recognition: A study in the operation of the reward system in science, *Am. Sociol. Rev.* **32**(3), 377-390.	0.048	14	0
s	Garcia *et al.* (2015), The author-editor game, *Scientometrics* **104**(1), 361-380.	0.048	0	8
1	Lee *et al.* (2013), Bias in peer review, *J. Assoc. Inf. Sci. Tec.* **64**(1), 2-17.	0.018	5	71
2	Zuckerman & Merton (1971), Patterns of evaluation in science: Institutionalisation, structure and functions of the referee system, *Minerva* **9**(1), 66-100.	0.016	73	2
3	Campanario (1998), Peer review for journals as it stands today: Part 1, *Sci. Commun.* **19**(3), 181-211.	0.013	23	35
4	Crane (1967), The gatekeepers of science: Some factors affecting the selection of articles for scientific journals, *Am. Sociol.* **2**(4), 195-201.	0.009	34	1
5	Campanario (1998), Peer review for journals as it stands today: Part 2, *Sci. Commun.* **19**(4), 277-306.	0.009	15	30
6	Gottfredson (1978), Evaluating psychological research reports: Dimensions, reliability, and correlates of quality judgments, *Am. Psychol.* **33**(10), 920-934.	0.008	26	2
7	Bornmann (2011), Scientific peer review, *Annu. Rev. Inform. Sci.* **45**(1), 197-245.	0.008	6	71
8	Bornmann (2012), The Hawthorne effect in journal peer review, *Scientometrics* **91**(3), 857-862.	0.007	1	20
9	Bornmann (2014), Do we still need peer review? An argument for change, *J. Assoc. Inf. Sci. Tec.* **65**(1), 209-213.	0.007	1	17
10	Merton (1968), The Matthew effect in science, *Science* **159**(3810), 56-63.	0.005	29	1

The core paths shown in Fig. 4B can be compared to the results obtained by Batagelj et al. (2017) using main path analysis. Different variants of main path analysis were used by Batagelj et al. (2017). Both using the original version of main path analysis (Hummon & Doreian, 1989) and using a more recent variant (Liu & Lu, 2012), the paths that were identified were rather lengthy, as can be seen in Figs. 9 and 10 in Batagelj et al. (2017). The shortest main paths included about 20 publications. This confirms the fundamental difference between intermediacy and main path analysis. Main path analysis tends to favor longer paths over shorter ones, whereas intermediacy has the opposite tendency.

Using the results presented in Table 2 and Fig. 4B, experts on the topic of peer review could discuss the historical development of the literature on this topic. Since our own expertise on the topic of peer review is limited, we refrain from providing an interpretation of the results.

Discussion

Citation networks provide valuable information for tracing the historical development of scientific knowledge. For this purpose, citation networks are usually analyzed using main path analysis (Hummon & Doreian, 1989). However, the idea of a main path is relatively poorly understood. The algorithmic definition of a main path is clear, but the underlying conceptual motivation remains somewhat obscure. As we have shown in this paper, main path analysis

has the tendency to favor longer paths over shorter ones. We consider this to be a counterintuitive property that lacks a convincing justification.

Intermediacy, introduced in this paper, offers an alternative to main path analysis. It provides a principled approach for identifying publications that appear to play a major role in the historical development from an older to a more recent publication. The older publication and the more recent one are referred to as the target and the source, respectively. Publications with a high intermediacy are important in connecting the source and the target publication in a citation network. As we have shown, intermediacy has two intuitively desirable properties, referred to as path addition and path contraction. Because of the path contraction property, intermediacy tends to favor shorter paths over longer ones. This is a fundamental difference with main path analysis. Intermediacy also has a free parameter that can be used to fine-tune its behavior. This parameter enables interpolation between two extremes. In one extreme, intermediacy identifies publications located on a shortest path between the source and the target publication. In the other extreme, it identifies publications located on the largest number of edge independent source-target paths.

We have also examined intermediacy in two case studies. In the first case study, intermediacy was used to trace historical developments at the interface between the community detection and the scientometric literature. This case study has shown that intermediacy yields results that appear sensible from the point of view of a domain expert. The second case study, in which intermediacy was applied to the literature on peer review, has provided an empirical illustration of the differences between intermediacy and main path analysis.

There are various directions for further research. First of all, a more extensive mathematical analysis of intermediacy can be carried out, possibly resulting in an axiomatic foundation for intermediacy. Intermediacy can also be generalized to weighted graphs. In a citation network, a citation link may for instance be weighed inversely proportional to the total number of incoming or outgoing citation links of a publication. Another way to generalize intermediacy is to allow for multiple sources and targets. The ideas underlying intermediacy may also be used to develop other types of indicators for graphs, such as an indicator of the connectedness of two nodes in a graph. In empirical analyses, intermediacy can be applied not only in citation networks of scientific publications, but for instance also in patent citation networks or in completely different types of networks, such as human mobility and migration networks, world trade networks, transportation networks, and passing networks in sports.

Source code

We calculate intermediacy using a Java implementation of a Monte Carlo algorithm. The source code is available at https://github.com/lovre/intermediacy (Šubelj, 2018).

Acknowledgements

We would like to thank Vladimir Batagelj for sharing the data used to study the literature on peer review (Batagelj et al., 2017). This work has been supported in part by the Slovenian Research Agency under the programs P2-0359 and P5-0168 and by the European Union COST Action number CA15109.

This paper is a shortened version of the full paper, available as a preprint from https://arxiv.org/abs/1812.08259.

References

Ball, M. (1980). Complexity of network reliability computations. *Networks*, *10*(2), 153–165.

Batagelj, V. (2003). Efficient algorithms for citation network analysis. *e-print arXiv:cs/0309023v1*, 1–27.

Batagelj, V., Doreian, P., Ferligoj, A., & Kejžar, N. (2014). *Understanding Large Temporal Networks and Spatial Networks*. Chichester: Wiley.

Batagelj, V., Ferligoj, A., & Squazzoni, F. (2017). The emergence of a field: A network analysis of research on peer review. *Scientometrics*, *113*(1), 503–532.

Chen, C. (2006). Citespace II: Detecting and visualizing emerging trends and transient patterns in scientific literature. *Journal of the American Society for Information Science and Technology*, *57*(3), 359-377.

Garfield, E. (2004). Historiographic mapping of knowledge domains literature. *Journal of Information Science*, *30*(2), 119–145.

Garfield, E., Pudovkin, A., & Istomin, V. (2003a). Mapping the output of topical searches in the Web of Knowledge and the case of Watson-Crick. *Information Technology and Libraries*, *22*(4), 183–187.

Garfield, E., Pudovkin, A., & Istomin, V. (2003b). Why do we need algorithmic historiography? *Journal of the American Society for Information Science and Technology*, *54*(5), 400–412.

Garfield, E., Sher, I., & Torpie, R. (1964, December). *The use of citation data in writing the history of science* (Tech. Rep. No. F49(638)-1256). The Institute for Scientific Information.

Gwak, J., & Sohn, S. (2018). A novel approach to explore patent development paths for subfield technologies. *Journal of the Association for Information Science and Technology*, *69*(3), 410–419.

Hummon, N., & Doreian, P. (1989). Connectivity in a citation network: The development of DNA theory. *Social Networks*, *11*(1), 39–63.

Kim, J., & Shin, J. (2018). Mapping extended technological trajectories: Integration of main path, derivative paths, and technology junctures. *Scientometrics*, *116*(3), 1439–1459.

Kuan, C.-H., Huang, M.-H., & Chen, D.-Z. (2018). Missing links: Timing characteristics and their implications for capturing contemporaneous technological developments. *Journal of Informetrics*, *12*(1), 259–270.

Liu, J., & Kuan, C.-H. (2016). A new approach for main path analysis: Decay in knowledge diffusion. *Journal of the Association for Information Science and Technology*, *67*(2), 465-476.

Liu, J., & Lu, L. (2012). An integrated approach for main path analysis: Development of the Hirsch index as an example. *Journal of the American Society for Information Science and Technology*, *63*(3), 528–542.

Lucio-Arias, D., & Leydesdorff, L. (2008). Main-path analysis and path-dependent transitions in HistCite-based historiograms. *Journal of the American Society for Information Science and Technology*, *59*(12), 1948–1962.

Marx, W., Bornmann, L., Barth, A., & Leydesdorff, L. (2014). Detecting the historical roots of research fields by reference publication year spectroscopy (RPYS). *Journal of the Association for Information Science and Technology*, *65*(4), 751–764.

Misra, K. (1970). An algorithm for the reliability evaluation of redundant networks. *IEEE Transactions on Reliability*, *R-19*(4), 146–151.

Moskowitz, F. (1958). The analysis of redundancy networks. *Transactions of the American Institute of Electrical Engineers*, *77*(5), 627–632.

Newman, M. (2018). *Networks* (2nd ed.). Oxford: Oxford University Press.

Page, L., & Perry, J. (1989). Reliability of directed networks using the factoring theorem. *IEEE Transactions on Reliability*, *38*(5), 556–562.

Park, H., & Magee, C. (2017). Tracing technological development trajectories: A genetic knowledge persistence-based main path approach. *PLoS ONE*, *12*(1), e0170895.

Thor, A., Marx, W., Leydesdorff, L., & Bornmannd, L. (2016). Introducing CitedReferencesExplorer (CRExplorer): A program for reference publication year spectroscopy with cited references standardization. *Journal of Informetrics*, *10*(2), 503–515.

Tu, Y.-N., & Hsu, S.-L. (2016). Constructing conceptual trajectory maps to trace the development of research fields. *Journal of the Association for Information Science and Technology*, *67*(8), 2016–2031.

van Eck, N., & Waltman, L. (2014). CitNetExplorer: A new software tool for analyzing and visualizing citation networks. *Journal of Informetrics*, *8*(4), 802–823.

van Eck, N., & Waltman, L. (2017). Accuracy of citation data in Web of Science and Scopus. In *Proceedings of the 16th International Conference on Scientometrics & Informetrics ISSI '17* (pp. 1087–1092). Wuhan, China.

Verspagen, B. (2007). Mapping technological trajectories as patent citation networks: A study on the history of fuel cell research. *Advances in Complex Systems*, *10*(1), 93–115.

Šubelj, L. (2018). *Intermediacy of publications.* http://dx.doi.org/10.5281/zenodo.1424365.

Yeo, W., Kim, S., Lee, J.-M., & Kang, J. (2014). Aggregative and stochastic model of main path identification: A case study on graphene. *Scientometrics*, *98*(1), 633–655.

Quantifying the long-term influence of scientific publications

Giovanni Colavizza[1], Massimo Franceschet[2], Vincent A. Traag[3] and Ludo Waltman[3]

[1] *g.colavizza@uva.nl*
University of Amsterdam
Centre for Science and Technology Studies (CWTS), Leiden University,
Willem Einthoven Building, Kolffpad 1, 2333 BN Leiden (The Netherlands)

[2] *massimo.franceschet@uniud.it*
Department of mathematics, computer science, and physics, University of Udine,
Via delle Scienze 206, 33100 Udine (Italy)

[3] *v.a.traag@cwts.leidenuniv.nl, waltmanlr@cwts.leidenuniv.nl*
Centre for Science and Technology Studies (CWTS), Leiden University,
Willem Einthoven Building, Kolffpad 1, 2333 BN Leiden (The Netherlands)

Abstract

We consider the long-term indirect influence of publications on subsequent publications. In particular, we are here interested in long-term scientific influence at the level of disciplines. We present a novel method to quantify the long-term scientific influence of publications, considering both direct and indirect, or higher-order citations. We apply this method to Web of Science data at the level of disciplines. Preliminary results for a specific operationalization of the notion of long-term scientific influence suggest that long-term influence is dominated by a few disciplines: *Astronomy and Astrophysics*, *Basic Life Sciences*, and *Physics and Material Science*.

Introduction

New knowledge builds on previous knowledge: this is a central tenet of science. A publication relies on previous publications and cites them to acknowledge this debt. Although citations acknowledge direct influences, the extent of the influence of a publication can go beyond these direct relations. In this paper, we present work in progress on quantifying how a publication relies on previous publications or, vice versa, how a publication influences ensuing publications. Our proposed method is somewhat similar to the well-known PageRank algorithm (Brin & Page, 1998; Franceschet, 2011; Waltman & Yan, 2014), but it is specifically focused on quantifying long-term scientific influence.

The study of the influence of previous publications on new ones rests at the core of scientometrics. The visualization and quantification of such relationships has been termed *algorithmic historiography* by Eugene Garfield (Garfield, Sher & Torpie, 1964; Garfield, Pudovkin & Istomin, 2003). A variety of tools have been developed for the purpose of facilitating such exploration (Chen, 2006; Marx et al. 2014; Thor et al. 2016; van Eck & Waltman, 2014). Our goal is to quantify scientific influence – thus give credit – beyond direct citations, and in particular to understand the long-term, indirect scientific influence of so-called fundamental research.

Method

Let $G = (V, E)$ be a citation network with V a set of n nodes and E a set of m directed edges. The nodes represent publications. If publication i cites publication j, then $(i, j) \in E$. Normally, G is a directed acyclic graph (DAG), because citations only go from more recent publications to older publications.[1] Let A be the adjacency matrix of citation network G, so that $A_{ij} = 1$ whenever i cites j, that is $(i, j) \in E$. Let d_i be the out-degree of node i, i.e. the number of

[1] There are some exceptions, but these can be removed so as to ensure that G is a DAG.

publications referenced by publication i within the citation network G. We recursively define the dependence of publication i on publication j as the mean dependence of publications referenced by i on j:

$$P_{ij} = \begin{cases} 1 \text{ if } i = j, \\ 0 \text{ if } i \neq j \text{ and } d_i = 0, \\ \frac{1}{d_i} \sum_k A_{ik} P_{kj} \text{ if } i \neq j \text{ and } d_i > 0. \end{cases}$$

We say that $\boldsymbol{P_{ij}}$ is the dependence of i on j, but it is also the influence of j on i. For example, $\boldsymbol{P_{ij}} = \frac{\mathbf{1}}{\boldsymbol{d_i}}$ if $\boldsymbol{i}$ directly cites $\boldsymbol{j}$ and no other paths exist between $\boldsymbol{i}$ and $\boldsymbol{j}$. In this case, $\boldsymbol{i}$ depends for $\frac{\mathbf{1}}{\boldsymbol{d_i}}$ on $\boldsymbol{j}$, and $\boldsymbol{j}$ has an influence of $\frac{\mathbf{1}}{\boldsymbol{d_i}}$ on $\boldsymbol{i}$.

The recursive equation always has a solution since the citation network G is acyclic. We can write this equation more compactly using matrix notation. Let D be a diagonal matrix such that $D_{ii} = \frac{1}{d_i}$ if $d_i > 0$ and $D_{ii} = 0$ otherwise. We can then write

$$P = DAP + I, \tag{1}$$

where I is the $n \times n$ identity matrix. We can solve for P and iteratively compute P because

$$P = (I - DA)^{-1} = \sum_{k=0}^{\infty} (DA)^k = \sum_{k=0}^{l} (DA)^k, \tag{2}$$

where $l \leq n - 1$ is the longest path in the citation network G. Matrix $(DA)^i$ gives the contribution to dependence of paths of length i in citation network G and since G is acyclic, $(DA)^k = 0$ for all $k > l$. Notice that $P_{ij} \neq 0$ if and only if there exists at least one path from i to j in citation network G.

Instead of looking at the individual dependence of publication i on publication j (i.e. the influence of publication j on i), we are interested in *disciplinary dependencies*. In particular, we are interested in:

1. the influence of a discipline on another discipline;
2. the influence of a discipline on all disciplines.

Let us denote by Q_{is} the extent to which publication i belongs to discipline s. Hence Q is an $n \times k$ matrix, where n is the number of publications and k is the number of disciplines. Publications may belong to multiple disciplines so that $Q_{is} > 0$ for possibly more than one discipline s. We always have $\sum_s Q_{is} = 1$ and $Q_{is} \geq 0$.

The *influence* F_{rs} of discipline r on discipline s is the sum of the influences of publications in r on publications in s, that is,

$$F = Q^T P^T Q,$$

where F is a $k \times k$ matrix. Note that we can compute $P^T Q$ without explicitly computing P, which would be computationally expensive. The *total influence* of disciplines on the whole network can simply be obtained by Fe, where e is the vector of all ones, and we can normalize by dividing by the total number of publications in each discipline $Q^T e$ to arrive at the *total influence per publication* in a discipline.

Data

We use data from the CWTS in-house version of the Web of Science, considering all publications between 2000 and 2016 included. We consider a total of 17,932,523 publications, and 190,550,206 citations among them – excluding 444,436 synchronous citations, which we discard to guarantee that G is a DAG.[2] The longest citation path in the dataset is of length 31 – equal to the maximum number of iterations needed for convergence.

Results

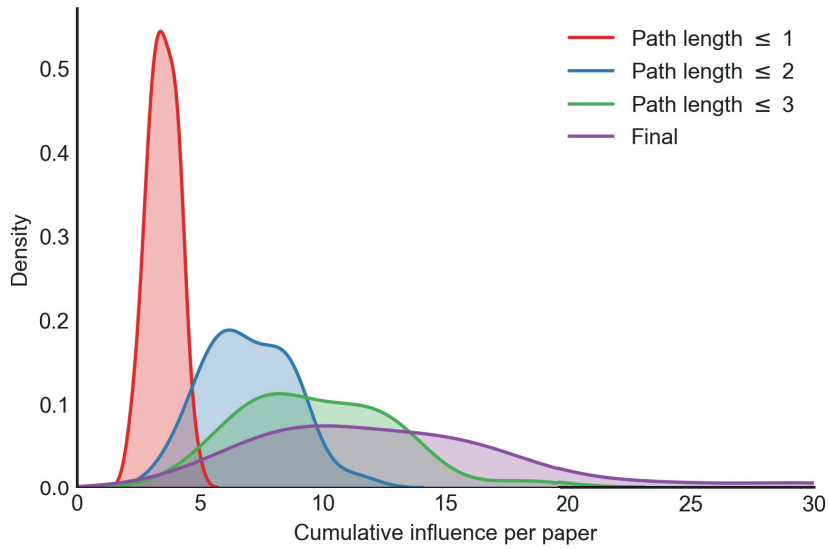


Figure 1. Distribution of influence per publication (year 2000) in 30 disciplines for different path lengths.

We examine the disciplinary influence of publications that appeared in the year 2000. We rely on a high-level aggregation of the journal-based classification of Web of Science, which represents 30 broad disciplines. As is shown in Eq. (2), we can separate the total influence F_{rs} into components of various path lengths. We pay particular attention to these path lengths and study how the influence differs for various path lengths.

Figure 1 shows the distribution of the cumulative influence per publication[3] of all 30 disciplines for various path lengths. If we consider only paths up to length 1 (i.e. only direct citations), we effectively calculate a type of source-normalized citation indicator. As is clear from Figure 1, the direct influence is quite concentrated around an influence per publication of about 3.5. When taking into consideration longer paths, differences between disciplines in the cumulative influence per publication increase. Eventually, the cumulative total influence per publication increases to 12.6 and the standard deviation increases from 0.6 to 5.2.

---

[2] A citation between two publications is discarded if the publication time (year and month) of the citing publication is the same, or older than the publication time of the cited publication.

[3] The influence of publications for path length ℓ is defined as $P_\ell = (DA)^\ell$, and similarly the total influence of disciplines on each other is defined as $F_\ell = Q^T {P_\ell}^T Q$. The cumulative influence is then defined as $P'_\ell = \sum_{k=0}^{\ell} P_k$ and similarly $F'_\ell = Q^T {P'_\ell}^T Q$. The total influence is then simply $F_\ell e$ and the cumulative total influence $F'_\ell e$, with corresponding influences per publication.

The total cumulative influence per publication is shown in Table 1. The discipline with the highest total cumulative influence per publication is *Basic Life Sciences*, followed by *Astrophysics and Astronomy*. This suggest that publications in these disciplines have a relatively large influence on scientific development, substantially beyond the direct impact as measured through direct citations.

Table 1 Total cumulative influence per publication (year 2000) for all disciplines.

Discipline	Influence	Discipline	Influence
BASIC LIFE SCIENCES	29.7	AGRICULTURE AND FOOD SCIENCE	11.1
ASTRONOMY AND ASTROPHYSICS	23.2	ELECTRICAL ENGINEERING AND TELECOMMUNICATION	11.1
BIOMEDICAL SCIENCES	18.2	COMPUTER SCIENCES	9.7
EARTH SCIENCES AND TECHNOLOGY	17.6	SOCIOLOGY AND ANTHROPOLOGY	9.7
CLINICAL MEDICINE	16.3	SOCIAL AND BEHAVIORAL SCIENCES, INTERDISCIPLINARY	9.6
PSYCHOLOGY	16.1	EDUCATIONAL SCIENCES	9.4
ENVIRONMENTAL SCIENCES AND TECHNOLOGY	15.8	MATHEMATICS	8.8
PHYSICS AND MATERIALS SCIENCE	15.8	INSTRUMENTS AND INSTRUMENTATION	8.7
BASIC MEDICAL SCIENCES	15.0	POLITICAL SCIENCE AND PUBLIC ADMINISTRATION	8.6
STATISTICAL SCIENCES	14.8	CIVIL ENGINEERING AND CONSTRUCTION	8.4
ECONOMICS AND BUSINESS	14.6	INFORMATION AND COMMUNICATION SCIENCES	7.9
CHEMISTRY AND CHEMICAL ENGINEERING	13.8	MECHANICAL ENGINEERING AND AEROSPACE	7.9
MANAGEMENT AND PLANNING	13.1	GENERAL AND INDUSTRIAL ENGINEERING	6.9
BIOLOGICAL SCIENCES	12.7	ENERGY SCIENCE AND TECHNOLOGY	6.4
HEALTH SCIENCES	12.3	LANGUAGE AND LINGUISTICS	5.2

Most of the influence of a discipline is concentrated in the first steps after publication, as shown in Figure 2. In some disciplines, this influence increases after the first step. For example, for *Basic Life Sciences*, the influence per publication in the first step is 4.8, while in the subsequent steps it increases to 6.5 and 7.0, only to decrease after the third step. For *Physics and Materials Science* this is less pronounced: the influence per publication in the first steps is 3.5, which then increases to 3.6 for the second step, but decreases already after the second step. For some other disciplines, the influence immediately decreases after the first step.

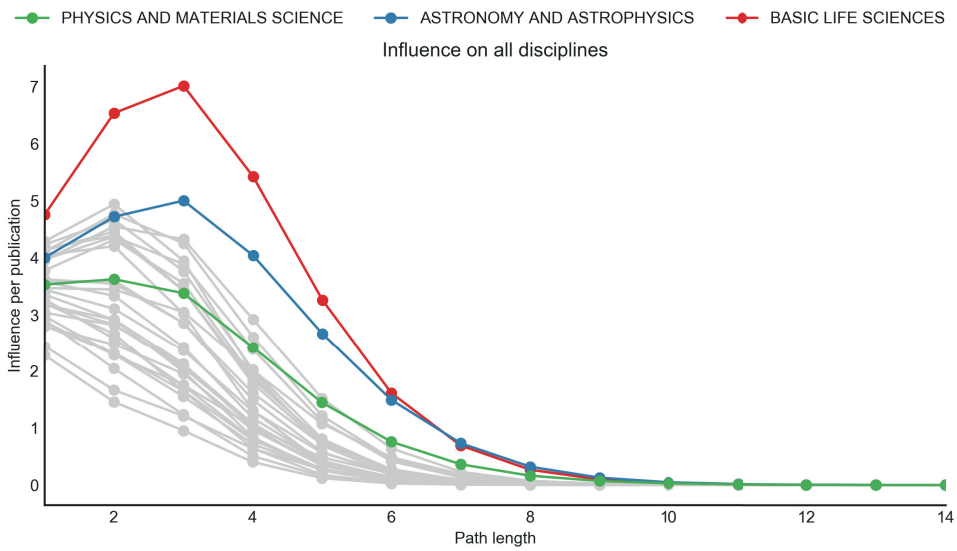


Figure 2. Influence per publication (year 2000), per path length.

Interestingly, there are substantial differences across disciplines in the influence per publication at later steps, which we interpret as long-term scientific influence. Figure 3 shows the relative influence per publication, for a given path length.[4] The long-term scientific influence seems to be dominated by three disciplines: *Astronomy and Astrophysics*, *Basic Life Sciences*, and *Physics and Material Science*.

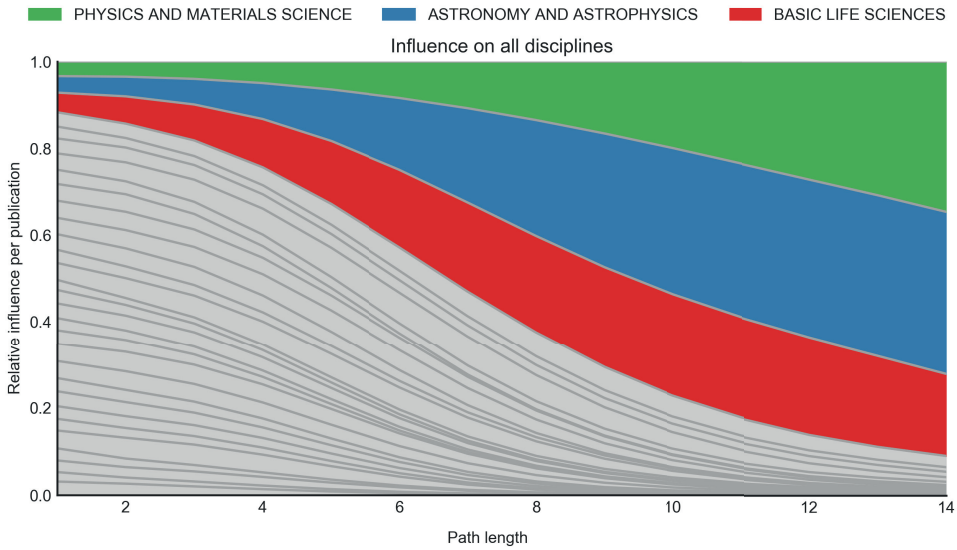


Figure 3. Relative influence per publication (year 2000), per path length.

[4] In other words, we calculate $F_\ell e$ relative to $e^T F_\ell$e.

Discussion and future work

We presented early work on quantifying the long-term scientific influence of individual publications and entire disciplines. We analyzed the results of our method using Web of Science data, considering all publications between 2000 and 2016, and studied the influence of publications that appeared in the year 2000 more in-depth. Using the definition of long-term scientific influence adopted in this paper, there are a few disciplines that turn out to be dominant in terms of long-term influence: *Astronomy and Astrophysics*, *Basic Life Sciences*, and *Physics and Material Science*. We suggest a few hypotheses that might explain our results, in view of exploring them in future work. Firstly, some fields might be more rapid in accumulating new literature which supersedes previous work and generates longer citation chains. Secondly, the citation culture in some fields might encourage direct citations as opposed to incremental ones, generating shorter chains and lower long-term influence, as captured by to our method. Thirdly, some results might need to be "translated" by "brokers" before impacting fields outside of their own, generating longer chains as a result.

Higher-order citations offer a promising approach for studying long-term scientific influence. However, we emphasize that there are different ways in which higher-order citations can be used to operationalize the idea of long-term scientific influence. Empirical results are likely to substantially depend on the specific operationalization that is adopted. The next step that we plan to take is to study, both theoretically and empirically, the differences between various ways in which citations can be used to quantify the long-term scientific influence.

References

Chen, C. (2006). Citespace II: Detecting and visualizing emerging trends and transient patterns in scientific literature. *Journal of the American Society for Information Science and Technology,* 57(3), 359-377.

Brin, S. & Page, L. (1998) The anatomy of a large-scale hypertextual web search engine. *Computer networks and ISDN systems*, 30(1-7), 107-117.

Franceschet, M. (2011) PageRank: Standing on the shoulders of giants. *Communications of the ACM.* 54(6), 92-101.

Garfield, E., Sher, I. & Torpie, R. (1964). *The use of citation data in writing the history of science*. The Institute for Scientific Information, Technical Report for Contract AF 49(638)-1256.

Garfield, E., Pudovkin, A. & Istomin, V. (2003). Why do we need algorithmic historiography? *Journal of the American Society for Information Science and Technology,* 54(5), 400-412.

Klavans, R. & Boyack, K. W. (2009). Toward a consensus map of science. *Journal of the American Society for Information Science and Technology*, 60(3), 455-476.

Marx, W., Bornmann, L., Barth, A. & Leydesdorff, L. (2014). Detecting the historical roots of research fields by reference publication year spectroscopy (RPYS). *Journal of the Association for Information Science and Technology,* 65(4), 751-764.

Thor, A., Marx, W., Leydesdorff, L. & Bornmann, L. (2016). Introducing CitedReferencesExplorer (CRExplorer): A program for reference publication year spectroscopy with cited references standardization. *Journal of Informetrics,* 10(2), 503-515.

van Eck, N.J. & Waltman, L. (2010). Software survey: VOSviewer, a computer program for bibliometric mapping. *Scientometrics*, *84*(2), 523-538.

van Eck, N.J. & Waltman, L. (2014). CitNetExplorer: A new software tool for analyzing and visualizing citation networks. *Journal of Informetrics,* 8(4), 802-823.

Waltman, L., & Yan, E. (2014). PageRank-related methods for analyzing citation networks. In Y. Ding, R. Rousseau, & D. Wolfram (Eds.), *Measuring scholarly impact: Methods and practice* (pp. 83-100). Springer.

Patent citations to scientific papers as early signs for predicting delayed recognition of scientific discoveries: a comparative study with instant recognition

Jian Du [1], Peixin Li [2], Robin Haunschild[3], Yinan Sun [4] and Xiaoli Tang*[5]

[1] *du.jian@imicams.ac.cn* [2] *li.peixin@imicams.ac.cn* [4] *sun.yinan@imicams.ac.cn* [5] *tang.xiaoli@imicams.ac.cn*
Institute of Medical Information & Library, Chinese Academy of Medical Sciences, Beijing, 100005, Beijing (China)

[3] *r.haunschild@fkf.mpg.de*
Max Planck Institute for Solid State Research, Heisenbergstraße 1, 70569 Stuttgart (Gemany)

Abstract

In this study, we investigate the extent to that patent citations to papers can serve as early signs for predicting delayed recognition using a comparative study with a control group, i.e., instant recognition papers. We identify the two opposite groups of papers by the Bcp measure, a parameter-free index for identifying papers which were recognized with delay (also called "sleeping beauties" in science). Combined with a typical case study, it appears that papers with delayed recognition show a stronger and longer technical impact than instant recognition papers. We provide indication that in the more recent years papers with delayed recognition are awakened more often and more earlier by a patent rather than by a scientific paper (also called "prince"). We also found that patent citations seem to play an important role to avoid instant recognition papers to level off or to become a so called "flash in the pan". It also appears that the sleeping beauties may firstly encounter negative citations and then patent citations and finally get widely recognized. In contrast to the two focus fields (biology and chemistry) for instant recognition papers, delayed recognition papers are rather evenly distributed in biology, chemistry, psychology, geology, materials science, and physics. We discovered several pairs of "science sleeping"-"technology inducing", such as biology-biotechnology/pharmaceuticals, chemistry-chemical engineering, psychology-computer/control technology, and physics-computer technology. We propose in further research to discover the potential ahead of time and transformative research by using citation delay analysis, patent & NPL analysis, and citation context analysis.

Introduction

According to our understanding of Kuhn's paradigm on the *Structure of Scientific Revolutions*, scientific knowledge proceeds incrementally (incremental research), occasionally punctuated by paradigm-shifting discoveries (transformative research) (Kuhn & Hawkins, 1963). In contrast to incremental research, which moves forward through the continuous, incremental accumulation of knowledge, transformative research drives science forward by radically changing our understanding of a concept, causing a paradigm shift, or opening new frontiers (Trevors, Pollack, Saier, & Masson, 2012). Prioritization of transformative research has become pervasive among funding agencies (Sen, 2017). Such research brings great rewards, but also carries great risks for funding agencies because transformative research projects are very hard to identify in their early stages. An on-going challenge lies in identifying transformative research projects at the time they are proposed. Although it is rarely possible to predict the transformative nature of research during the proposal stage, yet it is more predictable during the research process or even for a long time after the discovery. Further, transformative research should not be understood as just the opposite of incremental research. Actually, most transformative research began with incremental goals, and the transformative potential was recognized later (Gravem et al., 2017).

Premature discoveries and transformative research are crucial for the development of science, but they are often initially neglected or resisted by the scientific community and thus are subject to delayed recognition (Figure 1). In a report by the National Academies of Sciences,

Engineering, and Medicine in 2016, the committee reviewed five transformative areas of geographical research that have taken shape over the past 65 years to explore how transformative research has emerged. They found that transformative innovations can arise from older and long-ignored ideas (National Academies of Sciences & Medicine, 2016). Such ideas are often called "Sleeping Beauties" (SB) in science, one type of publications that goes unnoticed (or "sleeps") for a long time and then, almost suddenly, attracts a lot of attention (or "is awakened by a Prince") (A. F. J. Van Raan, 2004). This concept in terms of a citation curve is actually a quantitative description of "delayed recognition of scientific achievements", a phenomenon widely discussed in sociology of science (Hook, 2002). To the best of our knowledge sociologist Stephen Cole was the first to propose the notion of measuring delayed recognition in science using citations (Cole, 1970).

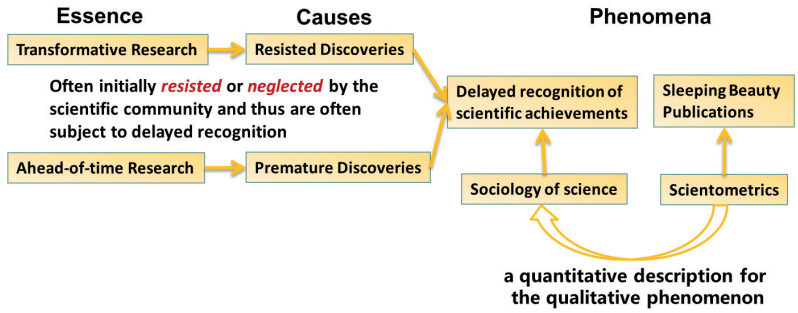


Figure 1. A schematic model of sleeping beauty and delayed recognition in science

Among the many papers written by colleagues on delayed recognition or SBs which have been discussed in our foregoing papers (Du & Wu, 2016, 2018), we concentrate on a few important recent developments and focus on the early identification of such type of publications and/or transformative research. Can we know in early stages if a research project looks promising or might lead to transformative research? The studies mentioned below may provide some insight into early signs of the awakening of SB publications or the recognition of premature discoveries and transformative research. It has been shown by Marx (2014) using the example of the paper by Shockley and Queisser (1961) that delayed recognition papers often start getting cited when a highly cited paper or another prominent author has drawn attention to them.

Dey, Roy, Chakraborty, and Ghosh (2017) analyzed the features of a given paper which may become an SB, and were the first to investigate the early identification of SBs in computer science. They developed a methodology to predict early, i.e., as soon as possible after the paper is published, whether a paper is likely to become an SB. By distinguishing two classes of papers—SBs and non-SBs—based on a set of features derived from each paper, they observed that the entropy of the number of fields from where the target paper has received citations is the most important feature. The more the paper has potential to attract attention from multiple fields, the higher the probability that it qualifies as an interdisciplinary paper that can become popular eventually. This observation is corroborated by Ke, Ferrara, Radicchi, and Flammini (2015). They found that in many cases, the awakening of SBs occurs when an application in a field outside of the SB's field is found, such as statistical methods that became useful in biology, chemistry, or physics.

Another important discovery is the correlation between the SBs and the scientific non-patent references in patents. One is perhaps more inclined to believe that SBs relate to more fundamental, basic, and less application-oriented work. But a surprising finding is that half of

the SBs in physics, chemistry, engineering, and computer science are application-oriented (A. F. J. Van Raan, 2015) and significantly more cited in patents than in scientific papers (A. F. J. van Raan, 2017). By investigating the time lag between the publication year of the SBs as scientific non-patent references (SNPR) in patents and their first citation in a patent, he found evidences that this time lag was becoming shorter in recent years (A. F. J. van Raan, 2017). In a very recent study, (A. F. J. van Raan & Winnink, 2018) investigated this further by using as recent as possible SBs cited in patents. Their observations suggest that, on average SBs are awakened increasingly earlier by a patent ("technological prince") rather than by a scholarly paper ("scientific prince") in the more recent years. It may suggest that SBs with technological importance are "discovered" earlier in an application-oriented context. Then, papers may be also cited in a scientific context because of the earlier recognized technological relevance. Thus, early recognized technological relevance may "prevent" papers from becoming an SB. Very recently, the scientific and technological impact of sleeping beauties in medical research fields was analyzed by Anthony F. J. van Raan and Winnink (2019). Du and Wu (2018) also found that 60% of the extreme SBs published in Science and Nature have been cited by patents, and the SB's first citation in terms of priority date (the earliest application date) in a patent usually appears to be earlier than the awakening year in the scientific context.

But studies by A. F. J. van Raan (2017) and A. F. J. van Raan and Winnink (2018) on patent citations to SBs are observational studies, not comparative studies. In their foregoing papers, they investigated a set of SBs with such thresholds as: (1) the average number of citations per year is at most one during 10 years after publication and, (2) the average number of citations per year is at least five during the next 10 years after 10 years of sleep. So, we can expect that the SBs have been cited at least 50 times. They identified 389 SBs for physics, 265 SBs for chemistry, and 367 SBs for engineering and computer science and found that 62 (16%), 92 (35%), and 108 (29%) of those SBs are also cited by patents. The possibility of SB papers being cited by patents is obviously higher than the proportion of all Web of Science (WoS) or MEDLINE covered publications cited by patents (about 4%). Ahmadpoor and Jones (2017) traced references from all 4.8 million patents issued by the U.S. Patent and Trademark Office (USPTO) from 1976 to 2015 to all 32 million journal articles published from 1945 to 2013 as indexed by the Web of Science. They found that 1.41 of 32 million (4.4%) WoS papers were cited by USPTO patents. This estimate is similar to a recent study where papers in MEDLINE rather than WoS were considered: a very small portion (4%) of MEDLINE papers published until 2013 are cited by USPTO patents (Ke, 2018). The percentage 4% is calculated based on all papers indexed by WoS/MEDLINE, including the papers which have never been cited. According to an investigation on the uncited papers in WoS (Van Noorden, 2017), the proportion of uncited papers levels off between five and ten years after publication in most disciplines, although the proportion is different in each discipline. Of all biomedical papers published in 2006, just 4% are uncited today; in chemistry, that number is about 8% and in physics it is close to 11%. In engineering and technology, the uncitedness rate is with 24% much higher than in the natural sciences. Thus, we can expect a dependence of the patent citation rate on scientific fields. Out of the 39 million research papers across all disciplines recorded in the Web of Science from 1900 to the end of 2015, about 21% haven't been cited, yet (Van Noorden, 2017). Since these papers have no scientific impact, they are likely to have no technical impact and thus will probably not be cited by patents.

In this study, we will answer two questions based on van Raan's work: (1) what is the extent to that patent citations can serve as early signs of delayed recognition using a comparative study with a control group, i.e., instant recognition papers？ Delayed recognition in science is a phenomenon where papers went unnoticed until they are re-discovered some years after publication. By contrast, instant recognition (also called "flashes in the pan") in science is a

phenomenon where papers received a lot of citations shortly after publication, but were ignored very quickly (Ye & Bornmann, 2018). (2) What is the pattern of the interaction in terms of citation relations between the sleeping science and the technology inducing its recognition? Based on our previous investigations on systematic identification of SBs and on their awaking mechanisms (Du & Wu, 2016, 2018), this contribution will further validate and detect early signs of the awakening of SBs. Our aim is to detect potential ahead-of-time discoveries or transformative research in order to shorten the time lag for original research to get recognized.

Data and Methods

Although being an SB sounds like a yes/no situation, it is clear that delayed recognition is not a clear-cut phenomenon (Rousseau, 2018). We are interested in detecting SBs in the hidden, under-cited publications with delayed impact. Note that our definition of "under-cited" is in contrast to "highly-cited". Currently, scholars in bibliometrics mostly focus on highly cited papers and ignore less highly cited or never cited ones. Research on SBs and flashes in the pan is valuable in turning scholars' attention to less highly or never cited papers which should have not been neglected. To characterize delayed recognition papers, it is necessary to compare them with instant recognition papers. We will turn an apparent yes/no question into a continuous phenomenon.

A parameter-free index for measuring the extent of delayed recognition

Based on the identification framework of the "beauty coefficient" (B) introduced by Ke et al. (2015) which takes the whole citation history of the publications concerned into account, Du and Wu (2018) substituted yearly citations in the beauty coefficient with yearly cumulative percentage of citations. The value of the modified beauty coefficient is denoted as Bcp. Bcp depends on the shape of the cumulative citation curve, especially when there is a cumulative citation burst in the whole life cycle, but not on the total number of citations of a given paper. Bcp works better than B in at least two aspects: (1) it "punishes" the situations when the SBs experienced early citations instead of continuous sleeping; (2) it allows comparisons of the extent of delayed citation impact of publications in different fields with different citation patterns.

In general, it is a sign of delayed recognition if a given paper's cumulative citation curve is concave (Bcp>0). Early citations are indicated by a convex cumulative citation curve (Bcp<0). The larger the Bcp value, the more delayed is the recognition of a paper in terms of the citation curve. The maximum value of the Bcp index is (n-1)/2, where n is the age of a given paper. This case occurs when the total number of citations is received in the last year. Hence, no citations are gathered in previous years. The smaller the Bcp's negative value, the more instant is the recognition of a paper. Just like the "top 1%" is usually used to select highly cited papers, we will also use "top 1%" versus "bottom 1%" for grouping delayed recognition and instant recognition papers.

In line with the earlier definition on the awakening year when the abrupt increase in the accumulation of citations of sleeping beauties occur, Du and Wu (2018) defined the "falling year" as the time when the abrupt decrease in the accumulation of citations of flash-in-the-pan papers occur. We will use this definition for "falling year" in our current study, too.

Delayed recognition (top 1%) versus instant recognition (bottom 1%) papers

The framework of Bcp was used to identify SBs published between 1970 and 2005 in Science and Nature. Citation data were included until the end of 2015. As we wish to have at least ten years of citation history after the latest publication year, 2005 is the most recent publication year included in our study. Articles with at least 200 citations, in total 20,000 publications

were included in the following analysis. These 20,000 papers were ordered by their Bcp value. We selected the top 1% (N=200) as delayed recognition papers and the bottom 1% (N=200) as instant recognition papers.

Patent-citing related indicators

Patent documents provide citations to earlier patents issued (prior art) and to non-patent literature (NPL), which includes peer-reviewed research and other published documents. Earlier patents may be cited by the inventor to demonstrate their difference from prior art or added by the examiner to limit the scope of the patent. Patent backward citations to NPL are considered stronger indicators of the impact of scientific research on technical invention than citations to patents (Roach & Cohen, 2013). So, using the patent backward citations to NPL, one can measure the technological impact of the scientific knowledge. We compare the extent to which the delayed recognition papers and the instant recognition papers show up as NPL in patents. In this study, the linkage between patents and NPLs was gathered by searching the platform lens.org, created by Cambia, a non-profit organization in Australia dedicated to facilitating innovation, and Queensland University of Technology. The platform lens.org has the world's patent information to most of the scholarly literature via collaborations with CrossRef and National Library of Medicine (Jefferson et al., 2018). "The Nature Index 2017 Innovation" published the top 200 institutions ranked by the Lens score, shedding new light on the impact academic research has had on innovation by examining how research articles are cited in third party patents[1]. We mainly focused on the following indicators.

1) Number of citing patent families: we group patent publications describing the same invention in "patent families" to prevent double counting when counting the number of patent citations to a given paper.
2) Interval of priority year between the earliest and the latest citing patent: by this measure, we can figure out the durability of patent citations to a given paper.

Fields of study

In order to compare the field of technology of papers with fields of study, we use the hierarchical fields of study from Microsoft Academic which are provided by a semantic algorithm on the paper basis. We appended the field of study from a local in-house database of Microsoft Academic to the top 1% and bottom 1% papers via the DOI and from Lens.org via PMID. Starting in August 2018, all scholarly papers cited by patents will have the information of field of study thanks to a partnership with Microsoft Academic[2]. Not all papers in Microsoft Academic database have a field of study attached to them but some papers have multiple fields of study at different levels. We found at least one field of study for 198 top 1% papers and for 196 bottom 1% papers with DOIs and/or PMIDs. For the rest of papers, we give the top level field of study based on their research areas reflected by title and/or abstract.

Results of a comparative study

Identifying the two opposite groups of papers by Bcp measure

Figure 2 shows citation curves of the first and the last paper ranked by Bcp and the distribution of citation percentiles for the two groups of papers. The awakening year for the most delayed recognition paper is 2004 (until the 33rd year after publication) and the falling year for the most instant recognition paper is 1977 (just in the 7th year after publication).

[1] https://www.nature.com/press_releases/nature-index-2017-innovation-supplement.html
[2] https://www.microsoft.com/en-us/research/project/academic/articles/sharpening-insights-into-the-innovation-landscape-with-a-new-approach-to-patents/

Many of the most delayed recognition papers are lowly cited, whereas many of the most instant recognition papers are highly cited. We can see that Bcp is not very dependent on the total number of citations of a given paper. It is appropriate for distinguishing those publications with delayed recognition from those with instant recognition although they are not so highly cited.

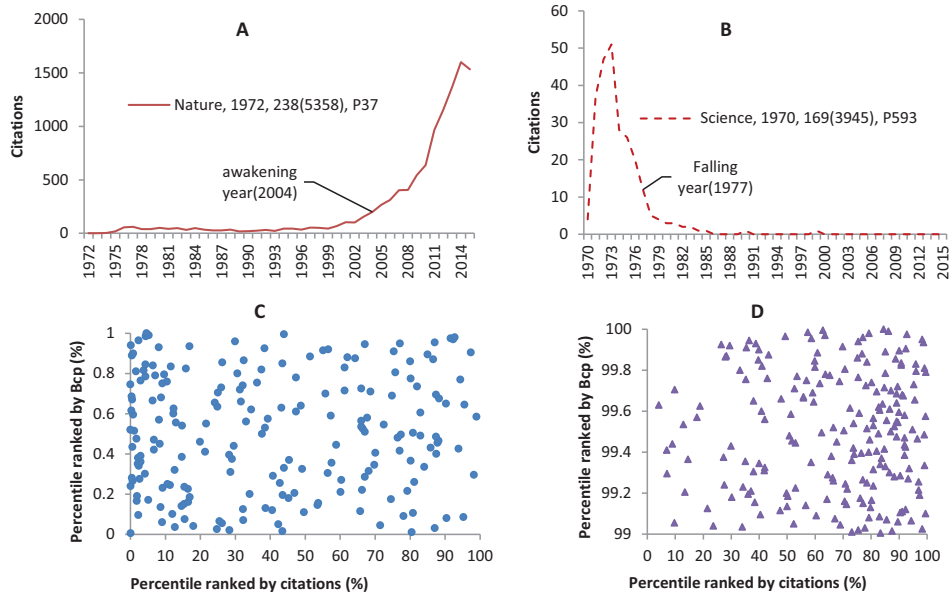


Figure 2. Citation curves of the first (A) and the last (B) paper ranked by Bcp and distribution of citation percentiles for top 1% (C) and bottom 1% (D) papers

Delayed recognition papers showing a stronger and longer technical impact than instant recognition papers

We find that about half of the 200 delayed recognition papers (DR-NPL) are cited by patents and about one-third of the 200 instant recognition papers (IR-NPL) are cited by patents.

Table 1. Extent to that the delayed recognition papers versus the instant recognition papers show up as NPLs in patents.

	top1%	*bottom 1%*
Total number of papers	200	200
Number of papers cited by patents	99	70
% of papers cited by patents	49.5	35.0
Total number of citing patent families	3988	543
Number of citing patent families per paper	40.3	7.8
Average durability of patent citations per paper	15.2	10.1

Note: Patent citations to non-patent literature accessed via lens.org by Dec 23, 2018.

Similar to citations given by publications, also the number of citations by patents is characterized by a skewed distribution. For instance, about one third of the DR-NPLs are cited by only 1 or 2 patents, and six are cited by more than 300 patents. In addition, about half of the IR-NPLs are cited by only 1 or 2 patents. One is cited by 120 patents, but the rest is cited by no more than 40 patents. In total, the 99 DR-NPLs are cited by 3988 patents, and the

70 IR-NPLs are cited by 543 patents. Delayed recognition articles were 5 times more cited by patents than instant recognition papers (on average 40.3 patent families versus 7.8 patent families), showing a stronger technical impact (Table 1). Next, we determined for the DR-NPLs and IR-NPLs the filing years of the earliest and of the latest citing patent until Dec 23, 2018. The difference between the filing years of the latest and the earliest patent indicates the durability of patent citations. The average durability per paper is 15 years for DR-NPLs and 10 years for IR-NPLs, showing a longer technical impact for delayed recognition papers.

Patent citations: earlier awakening, delayed recognition, and avoiding "flashes in the pan"

First, we compare the earliest patent citing year and the awakening year for the 99 delayed recognition NPLs, and find that for 70% (n=69) of the papers the first patent citing year is earlier than the awakening year; for 5% (n=5) of the papers the first patent citing year is the same as the awakening year; only 25% (n=25) of the papers are cited by patents after awakening. The difference between the filing year of the earliest citing patent and the awakening year, i.e., when the citations of the DR-NPL begin to abruptly increase define the time lag between the first citation by a patent and its "reviving". This time lag ranges from -19 to 29 years (average 6.7, SD=10.1). For example, for a Science article published in 1976 (10.1126/science.996549), the awakening year is 1994, and the year of the first patent citation is 2013. So, time lag between the year of the first patent citation and awakening year is -19. The average value relates to a long measuring period. In order to find out if there is a trend over time, we calculate the averages for successive, partly overlapping 5-year periods (Figure 3). In the case of the 99 DR-NPLs, these periods are 1970-1974, 1971-1975,…, and 1990-1994. Remarkably, the time lag is fluctuating in the earlier years but becomes rapidly shorter in the recent years. In other words, for the more recent DR-NPL, once it is cited by a patent, it will be awakened more quickly.

Afterwards, we analyze the time lag between the first patent citation and the falling year for instant recognition papers. The difference between the filing year of the earliest citing patent and the falling year, i.e., when the citations of the IR-NPL begin to abruptly decrease define the time lag between the first citation by a patent and its "perishing". Obviously, the time lag becomes rapidly longer in the measured period 1970-1983 (Figure 3). In the more recent years, even the instant recognition papers will be more likely to exhibit a long time window of citations once they are cited by a patent. Both observations suggest that, on average, in the more recent years, the delayed recognition papers are awakened increasingly earlier by a patent ("technological prince") rather than by a scholarly paper ("scientific prince"). Patent citations seem to play a more important role to avoid instant recognition papers to level off or become "flashes in the pan".

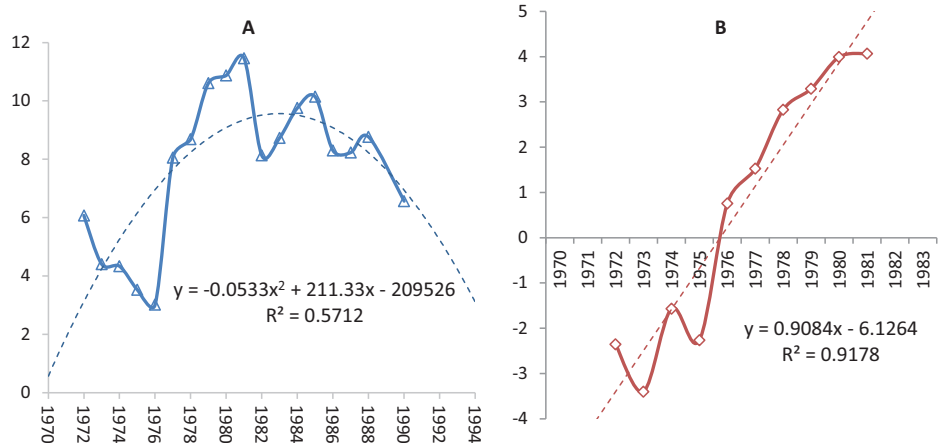


Figure 3. Time lag between the year of the first patent citation and the awakening year for DR-NPLs (A) and time lag between the first patent citation and the falling year for IR-NPLs (B).

Comparing the difference of science-technology interactions between DR and IR papers

The impact of discoveries may extend beyond the domain of science and may be crucial steps towards technological applications. It has been argued that technology-driven, or more specific, patent citations to papers, might be one of the awakening mechanisms for delayed recognition papers (Du, Sun, Zhang, & Tang, 2019). To reveal the whole picture of research fields for the scientific papers and the interactions with the technical focus of the citing patent families, we firstly match the field of study for each of the 200 delayed recognition and the other 200 instant recognition papers from Microsoft Academic, which determines the field of study based on machine learning parsing of all accessible text in a record. Microsoft Academic increases the power of semantic search by adding more fields of study[3] (from February 15, 2018). There are now 19 top-level fields of study, including Biology, Medicine, Geology, Chemistry, Psychology, Philosophy, Sociology, Engineering, Economics, Computer Science, Art, Physics, History, Political Science, Materials Science, Mathematics, Geography, Business, and Environmental Science. The Microsoft Academic data contain fields of study with a six-level hierarchy. Using the technology classification groups or WIPO concordance table[4], which links IPC symbols with 35 fields of technology we identified the fields of technology for each of the earliest citing patents in our two datasets. Afterwards, we map the interactions (by means of direct citations) between fields of study and fields of technology to figure out the different patterns for the two groups of papers.

There are 952 fields of study for the 200 delayed recognition papers, of which 55 (27.5%) are biology, followed by chemistry, psychology (n=25, 12.5%), geology (n=17, 8.5%), materials science (n=17, 8.5%), physics (n=16, 8%), and so on. However, there are 618 fields of study for the 200 instant recognition papers, of which almost 90% (n=180) are biology and nearly 10% (n=19) are chemistry. Figure 4 shows that delayed recognition papers in biology are mainly cited by patents in biotechnology and pharmaceuticals. Delayed recognition papers in chemistry are often cited by chemical engineering technology, biotechnology, and

[3] https://www.microsoft.com/en-us/research/lab/microsoft-research-asia/articles/microsoft-academic-increases-power-semantic-search-adding-fields-study/

[4] https://www.wipo.int/export/sites/www/ipstats/en/statistics/patents/pdf/wipo_ipc_technology.pdf

pharmaceuticals. Delayed recognition papers in materials science are mainly cited by patents in biotechnology and metallurgy materials. Delayed recognition papers in psychology are mainly cited by computer technology and control technology. Delayed recognition papers in physics are mainly cited by computer technology.

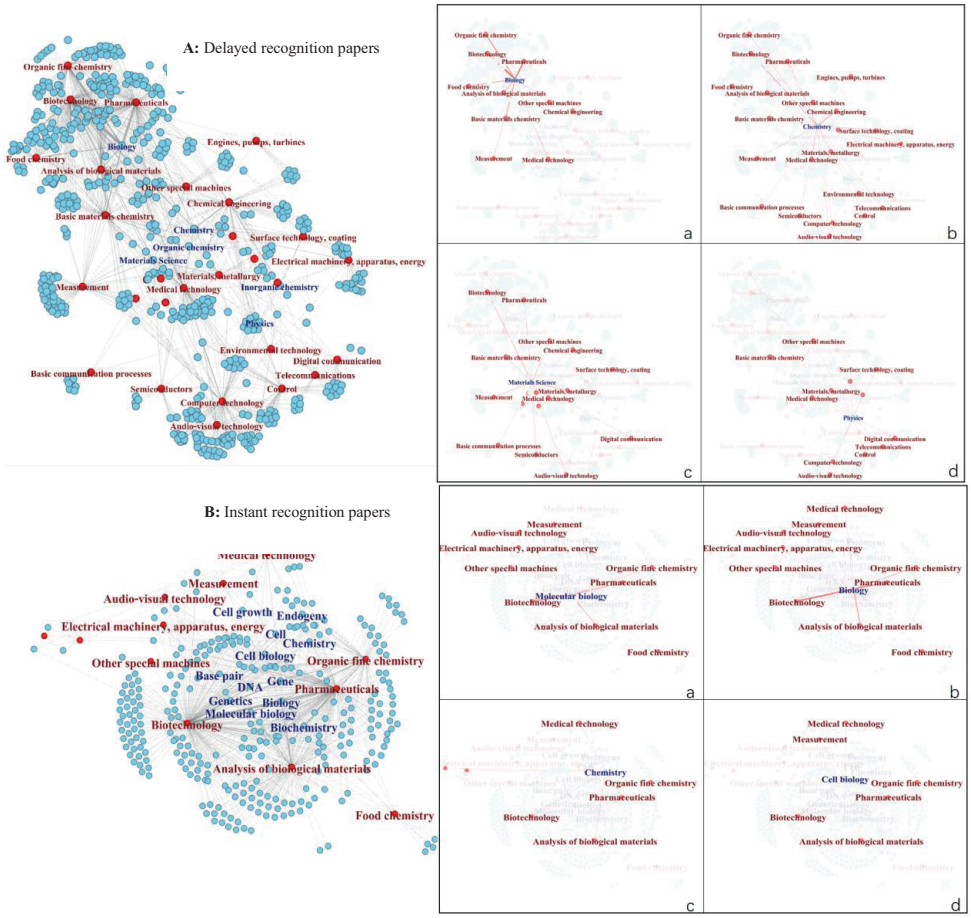


Figure 4. Interactions between fields of study (red color) of papers and their earliest citing patent families' fields of technology (blue color) for delayed recognition papers (A) and Instant recognition papers (B).

Discussion and Conclusion

Rousseau (2018) recently made an important observation that using citations to study delayed recognition is just a—convenient—operationalization of the concept, but that experts may agree on delayed recognition long before this is shown evidently by citations. He discovered a case in which an expert found delayed recognition several years before citation analysis could discover this phenomenon. This leads to the question: How good (or adequate) is citation analysis for detecting premature discoveries? To answer this question, we propose to prioritize the investigation into the lowly-cited papers instead of the most highly cited papers. Combining the frameworks of citation delay and beauty coefficient, we have proposed a parameter-free index known as Bcp index, for identifying under-cited SBs in science, which

may indicate possible breakthroughs in an early stage (Du & Wu, 2018). Note that our definition of “under-cited” is in contrast to “highly-cited”, from the perspective of the first generation of citations. By Bcp measure, we can distinguish delayed recognition from instant recognition papers.

Using articles published between 1970 and 2005 in the journals Science and Nature, we conducted a comparative study on delayed recognition with instant recognition papers. Combined with a case study, we found that delayed recognition papers show a stronger and longer technical impact than instant recognition papers. On average, in the more recent years the delayed recognition papers are cited increasingly earlier by a patent. Patent citations seem to play an important role to avoid instant recognition papers to level off or to become “flashes in the pan”. We provided further evidence to support the observation made by A. F. J. van Raan and Winnink (2018) that in the more recent years SBs are awakened increasingly earlier by a patent rather than by a scientific paper. This may suggest that early recognized technological relevance may “prevent” papers from becoming delayed recognition papers. It also appears that the sleeping beauties may firstly encounter negative citations, then patent citations, and finally get widely recognized.

We found that in contrast to the two focus fields (biology and chemistry) for instant recognition papers, delayed recognition papers are rather evenly distributed across biology, chemistry, psychology, geology, materials science, and physics. We also discovered several pairs of “science sleeping”-“technology inducing”, such as biology-biotechnology/pharmaceuticals, chemistry-chemical engineering, materials science-biotechnology/metallurgy materials, psychology-computer/control technology, and physics-computer technology.

The non-patent literature (NPL) cited by patents may provide insight into the awakening of delayed recognition publications, which may mean that the ahead-of-time discoveries get understood or the transformative potential of research is recognized. In a previous study, Du and Wu (2018) have discovered using citation context analysis that the extreme delayed recognition papers were all landmark publications of a specific research field, such as “the first report on …” or “the classic theory about …”. It appears that high quality publications tend to encounter delayed recognition and thus show delayed citation impact. One could identify transformative research through some text terms (such as "disagree", "contradict", "controversial", "inconsistent", "dispute", ...). In order to discern such potential transformative research, one could observe whether the relevant documents get early citation from patents or not. Much transformative research has influence on technology and invention and thus SBs tend to be cited by patents. In such cases, SBs with technological importance tend to be ‘discovered’ and ‘awakened’ earlier in an application-oriented context. Therefore, we propose to discover the potential ahead of time and transformative research by a combination of citation delay analysis, patent & NPL direct citation analysis, and citation context analysis.

Future research

Inspired by our investigations in this study, we propose to combine citation delay analysis with patent & NPL direct citation analysis to identify potential ahead of time and transformative research. The Bcp index proposed by Du and Wu (2018) can be used to identify those under-cited papers that are now happening to be at the sleeping-awakening interface. Afterwards, one could further identify those delayed recognition papers which are also cited by patents. Finally, one could map the structure of the older and long-ignored ideas at both the sleeping-awakening interface and science-technology interface. These ideas and research topics may be the potential origin of transformative research.

Further, inspired by National Institute of Health (NIH)'s Translational Science Search (http://tscience.nlm.nih.gov) and SciTech Strategies Inc.'s procedure for identifying discoveries in the biomedical sciences (Small, Tseng, & Patek, 2017), we argue that combining text mining based on authors' claims with citation context analysis from citers' comments, one may also discover potential transformative research. It may be possible to use text mining for identifying articles that are regarded by their authors as controversial (they challenge established dogma) or refutation (they disprove previously published data or hypotheses). The author's view can be compared with the citer's view by searching for specific terms (such as "disagree", "contradict", "contrast", "inconsistent", "dispute", ...) in the citation context. After a manual screening process to remove non-transformative research discoveries, it might be possible to provide a list of transformative research discoveries in the recent ten years from the perspectives of both author's claims and the community's comments.

Acknowledgments

This study was supported by the National Natural Science Foundation of China (Grant No. 71603280) and the Young Elite Scientists Sponsorship Program by China Association for Science and Technology (Grant No. 2017QNRC001) and CAMS Initiative for Innovative Medicine (CAMS-I2M-3-018). Data from Microsoft Academic (Sinha et al., 2015) (see also https://aka.ms/msracad) were shared with one of us (RH).

References

Ahmadpoor, M., & Jones, B. F. (2017). The dual frontier: Patented inventions and prior scientific advance. *Science, 357*(6351), 583-587.

Cole, S. (1970). Professional Standing and the Reception of Scientific Discoveries. *American Journal of Sociology, 76*(2), 286-306.

Dey, R., Roy, A., Chakraborty, T., & Ghosh, S. (2017). Sleeping beauties in Computer Science: characterization and early identification. *Scientometrics, 113*(3), 1645-1663. doi:10.1007/s11192-017-2543-3

Du, J., Sun, Y. N., Zhang, Y., & Tang, X. L. (2019). Characterizing and Detecting the early scientometric signs of the potential transformative research. *Bulletin of National Natural Science Foundation of China, 33*(150), 1-11.

Du, J., & Wu, Y. S. (2016). A Bibliometric Framework for Identifying “Princes” Who Wake up the “Sleeping Beauty” in Challenge-type Scientific Discoveries. *Journal of Data and Information Science, 1*(1), 50-68.

Du, J., & Wu, Y. S. (2018). A parameter-free index for identifying under-cited sleeping beauties in science. *Scientometrics, 116*(2), 959-971.

Gravem, S. A., Bachhuber, S. M., Fulton-Bennett, H. K., Randell, Z. H., Rickborn, A. J., Sullivan, J. M., & Menge, B. A. (2017). Transformative Research Is Not Easily Predicted. *Trends in Ecology and Evolution, 32*(11), 825-834.

Hook, E. B. (2002). *Prematurity in Scientific Discovery:On Resistance and Neglect*: University of California Press.

Jefferson, O. A., Jaffe, A., Ashton, D., Warren, B., Koellhofer, D., Dulleck, U., . . . Jefferson, R. A. (2018). Mapping the global influence of published research on industry and innovation. *Nature Biotechnology, 36*(1), 31-39.

Ke, Q. (2018). Comparing scientific and technological impact of biomedical research. *Journal of Informetrics, 12*(3), 706-717.

Ke, Q., Ferrara, E., Radicchi, F., & Flammini, A. (2015). Defining and identifying Sleeping Beauties in science. *Proceedings of the National Academy of Sciences of the United States of America, 112*(24), 7426-7431.

Kuhn, T. S., & Hawkins, D. (1963). The Structure of Scientific Revolutions. *American Journal of Physics, 31*(7), 554-555.

Marx, W. (2014). The Shockley-Queisser paper - A notable example of a scientific sleeping beauty. *Annalen der Physik, 526*(5-6), A41-A45.

National Academies of Sciences, E., & Medicine. (2016). *Fostering Transformative Research in the Geographical Sciences*. Washington, DC: The National Academies Press.
Roach, M., & Cohen, W. M. (2013). Lens or Prism? Patent citations as a measure of knowledge flows from public research. *Management Science, 59*(2), 504-525.
Rousseau, R. (2018). Delayed recognition: recent developments and a proposal to study this phenomenon as a fuzzy concept. *Journal of Data and Information Science, 3*(3), 1-13.
Sen, A. (2017). Island + Bridge: How transformative innovation is organized in the federal government. *Science and Public Policy, 44*(5), 707-721.
Shockley, W., & Queisser, H. J. (1961). DETAILED BALANCE LIMIT OF EFFICIENCY OF P-N JUNCTION SOLAR CELLS. *Journal of Applied Physics, 32*(3), 510-519.
Sinha, A., Shen, Z., Song, Y., Ma, H., Eide, D., Hsu, B.-J. P., & Wang, K. (2015). *An Overview of Microsoft Academic Service (MAS) and Applications.* Paper presented at the 24th International Conference on World Wide Web (WWW '15 Companion), Florence, Italy.
Small, H., Tseng, H., & Patek, M. (2017). Discovering discoveries: Identifying biomedical discoveries using citation contexts. *Journal of Informetrics, 11*(1), 46-62.
Trevors, J. T., Pollack, G. H., Saier, M. H., & Masson, L. (2012). Transformative research: Definitions, approaches and consequences. *Theory in Biosciences, 131*(2), 117-123.
Van Noorden, R. (2017). The science that's never been cited. *Nature, 552*(7684), 162-164.
Van Raan, A. F. J. (2004). Sleeping Beauties in science. *Scientometrics, 59*(3), 467-472.
Van Raan, A. F. J. (2015). Dormitory of physical and engineering sciences: Sleeping beauties may be sleeping innovations. *PLoS ONE, 10*(10).
van Raan, A. F. J. (2017). Sleeping beauties cited in patents: Is there also a dormitory of inventions? *Scientometrics, 110*(3), 1123-1156.
van Raan, A. F. J., & Winnink, J. J. (2018). Do younger Sleeping Beauties prefer a technological prince? *Scientometrics, 114*(2), 701-717.
van Raan, A. F. J., & Winnink, J. J. (2019). Sleeping Beauties in Medical Research: Technological Relevance, High Scientific Impact. Retrieved from https://arxiv.org/abs/1904.07658
Ye, F. Y., & Bornmann, L. (2018). "Smart girls" versus "sleeping beauties" in the sciences: The identification of instant and delayed recognition by using the citation angle. *Journal of the Association for Information Science and Technology, 69*(3), 359-367.

Synchronous scientific mobility and international collaboration: case of Russia

Denis Kosyakov[1] and Andrey Guskov[2]

[1] *kosyakov@spsl.nsc.ru*
State Public Scientific-Technological Library of the SB RAS, Voskhod Str 15, Novosibirsk (Russia)

[2] *guskov@spsl.nsc.ru*
State Public Scientific-Technological Library of the SB RAS, Voskhod Str 15, Novosibirsk (Russia)

Abstract

The phenomenon of multiple institutional affiliations of authors of research papers, or synchronous scientific mobility has emerged, partly in response to a scientometric quantitative approach to assessing the performance of scientific organizations in full accordance with Goodhart's law. Its wide distribution may distort the results of analysis of research collaborations based on bibliometric data. The article traces the influence of this phenomenon on the assessment of international collaborations of Russian researchers and organizations, evaluates the success of Russian programmes to attract leading foreign scientists. We show that in up to 20% of Russian international publications there are no authors with purely Russian affiliations. We identified at least 225 presumably invited researchers who have not published a single paper in collaboration with their Russian colleagues in 2014-2018.

Introduction

The importance of international collaborations in enhancing the performance and quality of research is widely recognized in the scientific world (J. Adams, 2013; J. D. Adams, Black, Clemmons, & Stephan, 2005; Aldieri, Kotsemir, & Vinci, 2018). The focus of integrating Russian science into the international community, strengthening international collaborations, including by attracting leading foreign scientists to Russian organizations, is secured by programme documents and decrees of the President of the Russian Federation. Reforms and restructuring of the Russian research and development sector have been taking place almost since the beginning of the 21st century, significant changes that have occurred over the past 5-6 years are described in (Block & Khvatova, 2017; Ivanov, Markusova, & Mindeli, 2016; Mindeli & Chernykh, 2016; Schiermeier, 2007). These changes, in aggregate, led to a significant increase in the productivity of Russian science (Kosyakov & Guskov, 2019a; Moed, Markusova, & Akoev, 2018; Shashnov & Kotsemir, 2018), and the transformation of its structure. An important driver of this growth has been the state scientific policy, focused on the use of quantitative performance indicators according to international databases, primarily the Web of Science. In this regard, Russian researchers have actively joined in the "publish or perish game", with this background, the interest in scientometric research in Russia has grown (A. Guskov, Kosyakov, & Selivanova, 2016).

In the contest of the rapid growth of the total publications number, the number of Russian publications in international collaborations grew much slower, and their share in the total flow has steadily decreased since 2007 despite all the efforts undertaken by the state to intensify international research cooperation (Shashnov & Kotsemir, 2018) (Fig. 1). The programmes to attract leading foreign scientists led to the emergence of foreign researchers in Russian universities and research institutes. Press releases and interviews with these scientists, such as Professor Roberto Morandotti, who took the position of a visiting professor at ITMO University[i], appeared in the news bulletins of the leading universities. One of the main objectives of such cooperation is the formation of permanent research team in Russian institution.Since foreign researchers are attracted to part-time and temporary positions they usually indicate several affiliations in their articles. However, some articles of these scientists

are published without the participation of Russian co-authors, while the Russian affiliation of this scientist is also indicated (Fig. 2).

There also occurs the opposite situation: when the Russian author indicates a foreign affiliation as additional, while all other authors indicate only Russian affiliations. Based on formal criteria, such publications are considered as written in international collaborations with Russian participation, although they are unlikely to be, since they can hardly be regarded as the result of the cooperation of researchers from Russia and other countries. We consider a research collaboration as a joint work of different researchers (Katz & Martin, 1997), and an international research collaboration as a joint work of researchers from different countries.

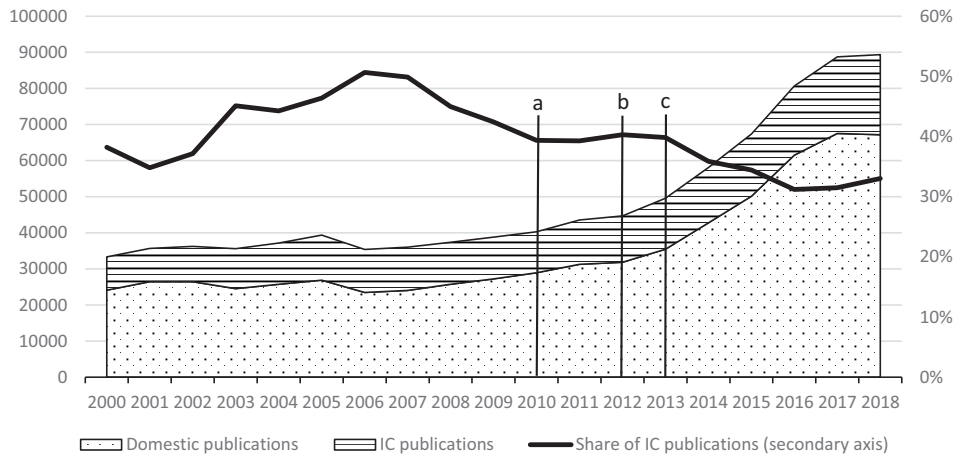
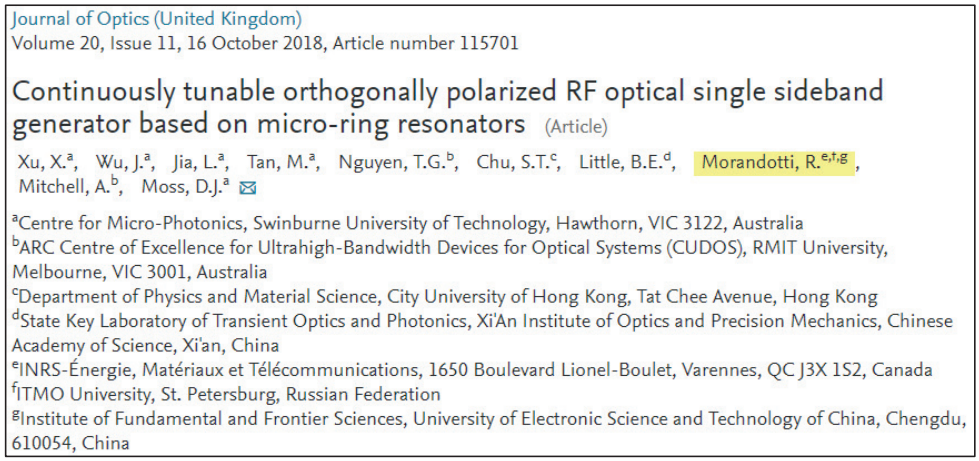


Fig. 1. Total number of Russian publications in Scopus, number of publications in international collaborations (IC) and share of IC publications in 2000-2018.

Milestones: a) Government Statement 220 on attracting foreign scientists, b) Decree 599 stated that the Russian share of research output (RO) has to reach 2.44% c) Reform of Russian State Academies of sciences, Government Statement 979 ordering to include bibliometric indicators in any research organization's evaluation



Journal of Optics (United Kingdom)
Volume 20, Issue 11, 16 October 2018, Article number 115701

Continuously tunable orthogonally polarized RF optical single sideband generator based on micro-ring resonators (Article)

Xu, X.[a], Wu, J.[a], Jia, L.[a], Tan, M.[a], Nguyen, T.G.[b], Chu, S.T.[c], Little, B.E.[d], Morandotti, R.[e,f,g], Mitchell, A.[b], Moss, D.J.[a]

[a]Centre for Micro-Photonics, Swinburne University of Technology, Hawthorn, VIC 3122, Australia
[b]ARC Centre of Excellence for Ultrahigh-Bandwidth Devices for Optical Systems (CUDOS), RMIT University, Melbourne, VIC 3001, Australia
[c]Department of Physics and Material Science, City University of Hong Kong, Tat Chee Avenue, Hong Kong
[d]State Key Laboratory of Transient Optics and Photonics, Xi'An Institute of Optics and Precision Mechanics, Chinese Academy of Science, Xi'an, China
[e]INRS-Énergie, Matériaux et Télécommunications, 1650 Boulevard Lionel-Boulet, Varennes, QC J3X 1S2, Canada
[f]ITMO University, St. Petersburg, Russian Federation
[g]Institute of Fundamental and Frontier Sciences, University of Electronic Science and Technology of China, Chengdu, 610054, China

Fig. 2. Screenshot from sample publication webpage in Scopus.

Research questions

The phenomenon of multiple affiliations itself has attracted the attention of researchers quite recently, primarily due to the proliferation of bibliometric methods for analysing scientific mobility and migration (Moed & Halevi, 2014; Robinson-Garcia et al., 2019). In the paper (Markova, Shmatko, & Katchanov, 2016) the phenomenon of multiple affiliations is analysed from the point of view of scientific mobility. The authors introduce the term "synchronous scientific mobility" (SSM), including international (SISM). Hotternrott and Lawson (Hottenrott & Lawson, 2017) assert, that the number of authors with multiple affiliations has at least doubled over the past few years in Germany, Japan and the UK in biology, chemistry, and engineering. Our calculations show that the share of Russian researchers indicating more than one affiliation in publications has also increased in recent years, and in 2017, the proportion of publications in which at least one of the authors indicated multiple affiliation reached 30% (Kosyakov & Guskov, 2019a).

Even the multiple co-authorship make it difficult to assess the contribution of individual institution in the national research output. Multiple affiliations of the authors make it much more complex task. Fractional count is one of the possible solution (Gauffriau, Larsen, Maye, Roulin-Perriard, & von Ins, 2007; Sivertsen, Rousseau, & Zhang, 2019). But at the moment the whole counting is widely adopted in adopted in Russia's practices of research performance assessment (Kosyakov & Guskov, 2019b), and strongly relies on affiliation data in the article's byline. We suppose that it has a significant impact on the high incidence of multiple affiliations in the last years.

Bhattacharjee (Bhattacharjee, 2011) draws attention to the fact that publications in which Saudi universities are indicated as an additional affiliation are based on studies conducted mainly in other places. Thus, part-time employment of the authors of these publications in those universities is a form of purchasing "academic prestige". Biagioli, Kenney, Martin, and Walsh (Biagioli, Kenney, Martin, & Walsh, 2018) mentioned the similar case in connection with the "Organizational gaming of the ranking". It can be noted that the "publish or perish" passion in the form of "gaming of the ranking" has shifted from the individual to the institutional and even the country level.

In the study of strategies aimed at increasing the publication activity of Russian universities-participants in the 5/100 project (A. E. Guskov, Kosyakov, & Selivanova, 2018), the authors have already paid attention to the type of publications in the collaboration between universities and, above all, research institutes of the Russian Academy of Sciences, in which, there is not a single author who has designated only university affiliation. We have selected these publications in a separate category and linked them to one of the dubious strategies for increasing publication activity.

Taking into account the number of publications where professor Morandotti is the only Russian-affiliated author and the ITMO affiliation is the second or third one we can assume that it is also a form of purchasing academic prestige. In this regard, it is interesting to what extent this practice is common in Russian scientific organizations, how many formally Russian publications are made without the participation of only Russian affiliated authors. In a more general sense, we are interested in assessing the influence of the multiple institutional affiliation phenomenon on the quantitative indicators of the international scientific collaboration of Russian researchers. It is also important to evaluate the performance of programmes to attract leading foreign scientists, identifying and attempting to categorize "Russian" authors indicating multiple international affiliations.

Data and methods

We use the Scopus database as a data source. This choice is due to several factors:

- Availability of an application programming interface (API) for downloading sufficiently detailed bibliographic records available in a standard subscription.
- Use of unique identifiers of authors and institutions to facilitate the implementation of the necessary calculations.
- A bibliographic record format which includes, in addition to the identifier, the country and city of affiliation.

Scopus warns that inaccurate identification of authors and affiliations is possible, and in this case, there may be duplicate profiles that may affect the calculations. We are trying to limit the influence of this factor, in some cases not taking into account the "long tail" in which duplicate profiles are concentrated. For the purposes of this study, we assume that the involvement of leading foreign scientists by Russian organizations is performed on a temporary basis, therefore, at least in their publications, they indicate both Russian and foreign affiliation, that is, they demonstrate synchronous international scientific mobility (SISM). The same applies to the facts of the temporary work of Russian researchers abroad. In this regard, for the purposes of analysis, SISM authors were selected. There is a possibility that some of the foreign scientists could immigrate to Russia on a permanent basis, breaking ties with the previous place of work and ceasing to indicate foreign affiliations in their publications. We cannot trace these authors with proposed approach.

Data acquisition and processing was carried out in the following sequence:

- The list of Russian publications for the years 2000-2018, obtained from the Scopus web interface by the request "AFFILCOUNTRY (Russian Federation) AND PUBYEAR AFT 1999", was downloaded in comma-separated value format using the Scopus web interface.
- According to this list, the bibliographic records of individual publications were downloaded, converted, and saved in the MongoDB database using the Powershell script using the Scopus Search API.
- Since the number of authors of single article is limited to 100 in the Scopus Search API answer, detailed bibliographic records of such publications were downloaded using the Scopus Abstract Retrieval API.
- From the two data collections obtained, the MongoDB script compiled a general list, the entries in which contain data about all publication authors and their affiliations in a form suitable for further processing.
- From this list, the MongoDB aggregated query obtained a list of publications in international collaborations (those in which affiliations from different countries were indicated) in the following classification:
    - If all publication authors have both Russian and foreign affiliations, this publication is of "Synchronous Publication (SP)" type.
    - If the publication does not belong to "SP" type and all publication authors have at least one Russian affiliation, this publication belongs to "Russian Publication (RP)" type.
    - If the publication does not belong to "SP" type and all authors have at least one foreign affiliation, this publication belongs to "Foreign Publication (FP)" type.
    - Other publications have both authors only with Russian affiliations and authors only with foreign affiliations. Such publications are of "International Publication (IP)" type.
- From the general list of publications, the MongoDB aggregated query also obtained a list of authors who indicated multiple affiliations, at least in part of their publications. For each of these authors, the number of publications was calculated, in which they indicated only foreign, only Russian affiliation and both Russian and foreign affiliations.

- For authors who indicated only Russian and foreign affiliations simultaneously with the use of the Scopus Search API, lists of their publications without Russian affiliations were obtained.
- Analysis of the publication history of the SISM authors allowed them to be classified according to the following principles:
    - The authors who in the part of publications indicated only Russian affiliations and never only foreign ones, were classified as "Russian Authors (RA)".
    - The authors who in the part of publications indicated only foreign affiliations and never only Russian ones, were classified as "Foreign Authors (FA)".
    - The authors who in part of their publications indicated only foreign and in another part only Russian affiliations were classified as "Migrated Authors (MA)".
    - The authors in all publications of which for the period under review both Russian and foreign affiliations were indicated were classified as "Synchronous Authors (SA)". Most of these authors have only one publication in the sample, which suggests that in this case we are dealing with a duplicate of the author's profile.
- A subset of Russian-affiliated publications of authors classified as "FA", and these authors' rankings by the number of publications for different periods of time allowed us to make assumptions about these authors' origin (in terms of Russian or not Russian) based on their names.
- Based on the affiliation data from publications of authors classified as "FA" for 2013-2018, a ranking of Russian organizations was also obtained by the number of invited foreign authors.

Results and discussion

The analysis performed yielded the following results.

Distribution of internationally collaborative papers of different assigned types

Table 1 presents data on the number of Russian publications in the international collaboration (IC publications) in accordance with the classification principles described above. It can be noted that the share of full-fledged collaborations (IP) decreased from 80% in 2009 to 70% by 2018 due to the increase in the number of publications with Russian co-authors, some of which indicated additional foreign affiliation (RP) and publications with foreign co-authors, some of which indicated additional Russian affiliation (FP). The share of the latter has grown to almost 20% of all Russian publications with international affiliation, which is also clearly seen in Fig. 3. As mentioned above, the number of IC publications more than doubled over the period under review, but with the background of a nearly threefold increase in the total number of publications, this led to a decrease in the share of IC publications in the total number.

Types of authors with multiple institutional affiliations from different countries

The classification of the authors according to the proposed principles gave the results presented in Table 2 and Fig. 4. Only those authors who, at least in the part of publications, have indicated both Russian and foreign affiliation are considered. The number of active authors, that is, having at least one publication in a given year, is indicated. It can be noted that the average annual number of invited foreign authors (FA) ranged within 600-300, reaching 322 in 2010, and since 2012, it has shown steady growth. We assume that this growth is associated with the effect of programmes to attract leading foreign scientists and stimulate publication activity that is reflected in the international databases.

Table 1. Number of Russian publications in Scopus database according to proposed classification of internationally collaborative papers

Years	*SP*	*RP*	*FP*	*IP*	*Total*
2000	409	277	1604	6935	9225
2001	398	271	1535	7012	9216
2002	411	291	1420	7700	9822
2003	348	303	1922	8505	11078
2004	370	287	1928	8817	11402
2005	388	403	1838	9857	12486
2006	398	408	1687	9404	11897
2007	329	358	1659	9648	11994
2008	349	388	1719	9135	11591
2009	387	399	1465	9308	11559
2010	392	404	1529	9060	11385
2011	414	435	1687	9739	12275
2012	415	601	1816	9998	12830
2013	412	730	1991	10989	14122
2014	468	829	2447	11595	15339
2015	511	1089	2871	12780	17251
2016	593	1365	3296	13934	19188
2017	569	1575	3694	15414	21252
2018	611	1842	4331	15411	22195

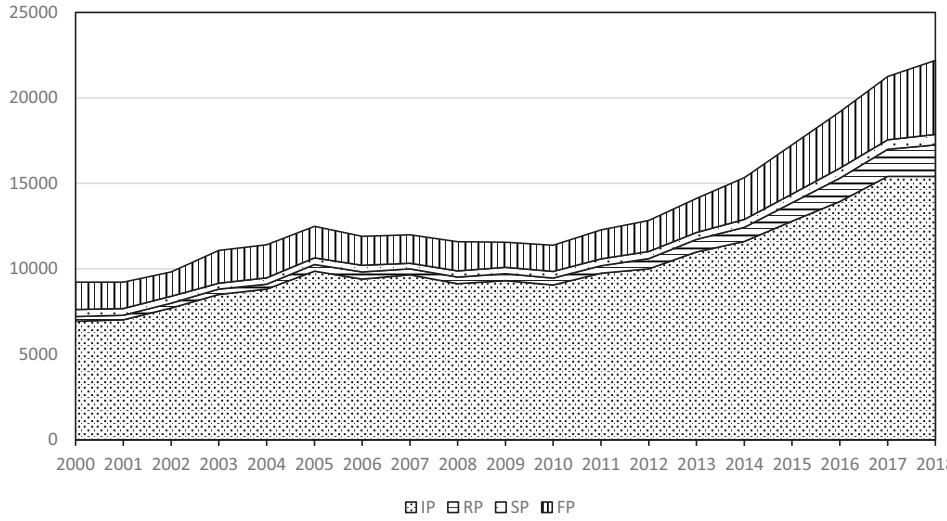


Fig. 3. Internationally collaborative Russian publications according to proposed classification

Table 2. Number of SISM authors by category and by year according to proposed classification

Years	*FA*	*MA*	*RA*	*SA*	*Total*
2000	599	2420	3192	20	6231
2001	553	2521	3316	19	6409
2002	655	2571	3530	13	6769
2003	390	2438	3505	14	6347
2004	344	2465	3678	24	6511
2005	452	2511	3912	22	6897
2006	427	2551	3900	10	6888
2007	544	2573	3991	18	7126
2008	296	2633	4124	18	7071
2009	364	2644	4196	10	7214
2010	322	2697	4367	6	7392
2011	448	2834	4470	15	7767
2012	391	2866	4510	24	7791
2013	436	2938	4744	19	8137
2014	606	3041	4945	38	8630
2015	777	3146	5155	45	9123
2016	1072	3236	5202	82	9592
2017	1146	3208	5266	94	9714
2018	1186	3077	5043	131	9437

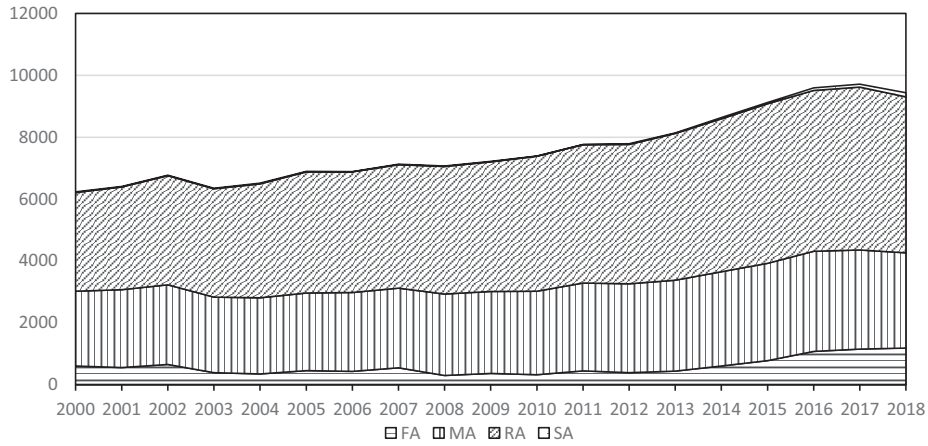


Fig. 4. Number of SISM authors by category and by year according to proposed classification

Number of invited researchers and host institutions

Publications for 2014–2018 were chosen as the basis for the analysis, since during this period all factors related to the state scientific policy to attract leading foreign authors were fully operational. In order to reduce errors due to inaccuracy of data in Scopus, authors were selected with more than one publication over the entire period 2000-2018. We counted 2,253 such authors that had publications with Russian affiliations in 2014-2018, while the total number of these authors for the entire period 2000-2018 was 7,284.

The most active organizations in attracting foreign authors are shown in Table 3.

Table 3. Top-20 of Russian institutions by number of invited foreign researchers in 2014-2018

Institution	*City*	*Number of researchers*
Joint Institute for Nuclear Research	Dubna	185
Lomonosov Moscow State University	Moscow	168
National Research University Higher School of Economics	Moscow	161
Novosibirsk State University	Novosibirsk	152
ITMO University	Saint Petersburg	152
National Research Nuclear University MEPhI	Moscow	129
Saint Petersburg State University	Saint Petersburg	122
Tomsk State University	Tomsk	115
Kazan Federal University	Kazan	109
Moscow Institute of Physics and Technology	Moscow	100
Tomsk Polytechnic University	Tomsk	88
Peter the Great St. Petersburg Polytechnic University	Saint Petersburg	78
RUDN University	Moscow	74
National University of Science & Technology (MISIS)	Moscow	74
Sechenov First Moscow State Medical University	Moscow	74
P.N. Lebedev Physical Institute of RAS	Moscow	59
Skolkovo Institute of Science and Technology	Moscow	55
Petersburg Nuclear Physics Institute (PNPI)	Gatchina	53
Ural Federal University	Ekaterinburg	44
Lobachevsky State University of Nizhni Novgorod	Nizhni Novgorod	43

Examination of the names of FA authors entering the top by publication activity for different periods of time allowed us to suggest that at the beginning of the period (before 2010), the majority of these authors were representatives of the Russian diaspora, scientists who emigrated from Russia in the 1990-s, but retained scientific ties with their alma mater organizations. In the latter period (2014-2018) there is a significant number of foreign authors, probably attracted under the relevant programmes in the top of the authors' ranking by publication activity. Many of these authors are also characterized by a significant number of FP publications, which are Russian only due to their additional Russian affiliation. The average proportion of such publications for FA authors is 45%, but for Professor Morandotti mentioned above, it is equal to 94%. For this period, only 4 out of 71 articles were published by Professor Morandotti in collaboration with his colleagues from ITMO University. More than two hundred authors (225) have not published any work at all in collaboration with their Russian colleagues, confining themselves to attributing additional Russian affiliation, which represents a significant 10% of the total number of FA authors for this period. A typical example is Rafael Luque, a chemist from Universidad de Córdoba, who mentioned Peoples Friendship University of Russia (RUDN University) as an additional affiliation in 45 publications for 2018. None of these publications contains co-authors not only from RUDN University but also from Russia in general.

Conclusion

The analysis shows that the practice of multiple affiliations can significantly affect both the general indicators of the publication activity of organizations and countries, and the assessment of international collaborations. Up to 20% of Russian publications related to international collaborations are attributed to Russia only because some authors of these works for some reason indicated the Russian organization as an additional one. Presumably, at least in part this

practice can be interpreted as the purchase of academic prestige. Where there are buyers, there are sellers. In fact, a community has emerged constituting researchers who are ready to peddle academic prestige – to assign additional affiliations for remuneration in one or another form. At least 225 presumably invited foreign researchers from Russian institutions have not published a single paper in collaboration with their Russian colleagues in 2014-2018. Of course, one cannot state the existence of scientific misconduct based only on bibliometric metadata, but we suppose that these cases require close study.

A significant number of papers are devoted to the discussion of ethical issues of multiple authors of scientific publications (see, for example, (Teixeira da Silva & Dobránszki, 2016). In their recent paper in Nature Ioannidis, Klavans, and Boyack (Ioannidis, Klavans, & Boyack, 2018) stated that "Loose definitions of authorship, and an unfortunate tendency to reduce assessments to counting papers, muddy how credit is assigned". Affiliations indicated by the article authors are currently used in a wide range of scientometric tasks, including in the calculation of different rankings, in the research assessment of institutions, even in informal research performance races between countries. As in the case of authorship, it is implied that the affiliations indicated in the scientific article are related to the contribution of relevant institutions to the research, expressed in the provision of funding, workplace, instrumentation base, samples, reagents, source data, etc. Thus, there is a lack of ethical principles widely accepted by the scientific community that determine the legitimacy of specifying a particular affiliation.

Acknowledgments

This study was supported by the Russian Foundation for Basic Research (Grant No 18-011-00797 A) and the Ministry of Science and Higher Education of the Russian Federation (Project No 0334-2019-0006).

References

Adams, J. (2013). The fourth age of research. *Nature*, *497*(7451), 557–560. https://doi.org/10.1038/497557a

Adams, J. D., Black, G. C., Clemmons, J. R., & Stephan, P. E. (2005). Scientific teams and institutional collaborations: Evidence from U.S. universities, 1981-1999. *Research Policy*, *34*(3), 259–285. https://doi.org/10.1016/j.respol.2005.01.014

Aldieri, L., Kotsemir, M., & Vinci, C. P. (2018). The impact of research collaboration on academic performance: An empirical analysis for some European countries. *Socio-Economic Planning Sciences*, *62*, 13–30. https://doi.org/10.1016/j.seps.2017.05.003

Bhattacharjee, Y. (2011). Saudi universities offer cash in exchange for academic prestige. *Science*, *334*(6061), 1344–1345. https://doi.org/10.1126/science.334.6061.1344

Biagioli, M., Kenney, M., Martin, B. R., & Walsh, J. P. (2018). Academic misconduct, misrepresentation and gaming: A reassessment. *Research Policy*, *48*(November 2018), 401–413. https://doi.org/10.1016/j.respol.2018.10.025

Block, M., & Khvatova, T. (2017). University transformation: explaining policy making and trends in higher education in Russia. *Journal of Management Development*, *36*(6), 761–779. https://doi.org/10.1108/JMD-01-2016-0020

Gauffriau, M., Larsen, P. O., Maye, I., Roulin-Perriard, A., & von Ins, M. (2007). Publication, cooperation and productivity measures in scientific research. *Scientometrics*, *73*(2), 175–214. https://doi.org/10.1007/s11192-007-1800-2

Guskov, A. E., Kosyakov, D. V., & Selivanova, I. V. (2018). Boosting research productivity in top Russian universities: the circumstances of breakthrough. *Scientometrics*. https://doi.org/10.1007/s11192-018-2890-8

Guskov, A., Kosyakov, D., & Selivanova, I. (2016). Scientometric research in Russia: impact of science policy changes. *Scientometrics*, *107*(1). https://doi.org/10.1007/s11192-016-1876-7

Hottenrott, H., & Lawson, C. (2017). A first look at multiple institutional affiliations: a study of authors

in Germany, Japan and the UK. *Scientometrics*, *111*(1), 285–295. https://doi.org/10.1007/s11192-017-2257-6

Ioannidis, J. P. A., Klavans, R., & Boyack, K. W. (2018). Thousands of scientists publish a paper every five days. *Nature*, *561*(7722), 167–169. https://doi.org/10.1038/d41586-018-06185-8

Ivanov, V. V, Markusova, V. A., & Mindeli, L. E. (2016). Government investments and the publishing activity of higher educational institutions: Bibliometric analysis. *Herald of the Russian Academy of Sciences*, *86*(4), 314–321. https://doi.org/10.1134/S1019331616040031

Katz, J. S., & Martin, B. R. (1997). What is research collaboration? *Research Policy*, *26*(1), 1–18. https://doi.org/10.1016/S0048-7333(96)00917-1

Kosyakov, D., & Guskov, A. (2019a). Impact of national science policy on academic migration and research productivity in Russia. *Procedia Computer Science*, *146*, 60–71. https://doi.org/10.1016/j.procs.2019.01.080

Kosyakov, D., & Guskov, A. (2019b). Research assessment and evaluation in Russian fundamental science. *Procedia Computer Science*, *146*, 11–19. https://doi.org/10.1016/j.procs.2019.01.072

Markova, Y. V., Shmatko, N. A., & Katchanov, Y. L. (2016). Synchronous international scientific mobility in the space of affiliations: evidence from Russia. *SpringerPlus*, *5*(1), 480. https://doi.org/10.1186/s40064-016-2127-3

Mindeli, L. E., & Chernykh, S. I. (2016). Funding of basic research in Russia: Modern realities and forecasts. *Studies on Russian Economic Development*, *27*(3), 318–325. https://doi.org/10.1134/S1075700716030114

Moed, H. F., & Halevi, G. (2014). A bibliometric approach to tracking international scientific migration. *Scientometrics*, *101*(3), 1987–2001. https://doi.org/10.1007/s11192-014-1307-6

Moed, H. F., Markusova, V., & Akoev, M. (2018). Trends in Russian research output indexed in Scopus and Web of Science. *Scientometrics*, *116*(2), 1153–1180. https://doi.org/10.1007/s11192-018-2769-8

Robinson-Garcia, N., Sugimoto, C. R., Murray, D., Yegros-Yegros, A., Larivière, V., & Costas, R. (2019). The many faces of mobility: Using bibliometric data to measure the movement of scientists. *Journal of Informetrics*, *13*(1), 50–63. https://doi.org/10.1016/j.joi.2018.11.002

Schiermeier, Q. (2007). Russian science: The battle for Russia's brains. *Nature*, *449*(7162), 524–527. https://doi.org/10.1038/449524a

Shashnov, S., & Kotsemir, M. (2018). *Research landscape of the BRICS countries: current trends in research output, thematic structures of publications, and the relative influence of partners*. *Scientometrics* (Vol. 7). Springer International Publishing. https://doi.org/10.1007/s11192-018-2883-7

Sivertsen, G., Rousseau, R., & Zhang, L. (2019). Measuring scientific contributions with modified fractional counting. *Journal of Informetrics*, *13*(2), 679–694. https://doi.org/10.1016/j.joi.2019.03.010

Teixeira da Silva, J. A., & Dobránszki, J. (2016). Multiple Authorship in Scientific Manuscripts: Ethical Challenges, Ghost and Guest/Gift Authorship, and the Cultural/Disciplinary Perspective. *Science and Engineering Ethics*, *22*(5), 1457–1472. https://doi.org/10.1007/s11948-015-9716-3

[i] Interview with prof. Morandotti - http://news.ifmo.ru/ru/science/photonics/news/6741/ (accessed 20-01-2018, employee profile at the site of ITMO University - http://edu.ifmo.ru/teacher/231253/ (accessed 20-01-2018).

A Deep-Learning Approach to Determine the Dependency between Two Subject Types in the Web of Science

Frederick Kin Hing Phoa[1], Hsin-Yi Lai[2], Livia Lin-Hsuan Chang[3] and Keisuke Honda[4]

[1] *fredphoa@stat.sinica.edu.tw*
Institute of Statistical Science, Academia Sinica, Taipei City 11529 (Taiwan)

[2] *nanaba85224@gmail.com*
Institute of Statistics, National Chiao Tung University, Hsinchu 30010 (Taiwan)

[3] *livia@ism.ac.jp*
SOKENDAI (The Graduate University for Advanced Studies), Tokyo 190-8562 (Japan)

[4] *khonda@ism.ac.jp*
Institute of Statistical Mathematics, Tokyo 190-8562 (Japan)

Abstract
Traditional wisdom always suggests that some subjects have strong relationships, while others are almost mutual independent. However, there is a lack of quantitative approach to formulate these relationships in a systematic and unbiased way. In this work, we train a classification machine via deep learning to determine whether two subject types are independent based on the citation information from the Web of Science database. This machine not only achieves very high accuracy in estimating the dependency among subject types in the database, but also is able to predict the dependency when one or both subject types do not exist in the database.
Keywords: Deep Learning, Multilayer Perceptron, Classification, Web of Science, Dependency

Introduction

It is a common sense that some subjects have strong relationships but others are mutual independent. For example, one expects that applied mathematics and physics should be very close to one another due to its nature and subject evolution, and both subjects may relatively have little relationship with literature. Even though the subject relationship can be described roughly via experience, expert knowledge or any subjective perspectives, it is a lack of a systematic and quantitative way to characterize this important quantity.

In the regime of big data and advanced computing power, network analysis helps to reveal the relationships between two objects, which can be species in ecology (Yu et al. 2014), neurons in human nervous systems (Ascoli 2002), electric components in electric circuit networks (Bakshi and Bakshi 2008), web users in the Internet (Phoa and Sanchez 2013), human beings in social communities (Wang and Phoa 2015), and many others. Recently, the citation network receive much attentions to researchers in scientific communities, especially in the institutional researches, as it provides a scientific method to evaluate the research performance of an author or an institute. For example, Chang, Phoa and Nakano (2018) studied the article citations by considering research articles as nodes and the citations between articles as edges. The citation networks can be obtained from a transformation of large-scale citation databases like the Web of Science (WoS), Scopus, Google Scholars, or others.

Our analysis is based on the data from the WoS. There have been ongoing projects of analysing WoS at article-, author-, institute- and many other levels. This work particularly focuses on the analysis on the journal and subject levels. Up to date, there are 20327 journals, which are classified into 275 distinct subjects and 3120 subject types, in the WoS database. The purpose of this work is to introduce a systematic quantity modified from the standard pointwise mutual information (PMI) to describe the relationship between two subject types, and to propose an efficient and feasible approach to obtain this quantity. The rest of the paper is

organized as follows. Section 2 provides the definition of PMI. Section 3 introduces the details of the classification step via deep learning. Section 4 shows two results using different number of layers. The discussion on future researches is given in the last section.

Pointwise Mutual Information (PMI)

The pointwise mutual information (PMI) is a measure of association used in information theory and statistics. According to Church and Hanks (1990), the PMI of a pair of outcomes x and y belonging to discrete random variables X and Y quantifies the discrepancy between the probability of their coincidence given their joint distribution and their individual distributions, assuming independence. This measure has also been used for finding co-occurrences of words in a message (Roder, Both and Hinneburg 2015).

We implement the PMI in the WoS framework below. First, we define a subject type as a word string that consists of the names of related but individual subjects identified in the journals of the WoS. There are 3120 subject types and each of them can be a pure individual subject or a complex subject combination that consists of N individual subjects in the form of $S_1 + S_2 + \cdots + S_N$. Let X and Y be the "cite" and "cited" subject types respectively. In the language of networks, we consider a direct edge $X \to Y$ on the number of citations of y that come from x. Then the PMI of $X \to Y$ is defined as

$$PMI(X \to Y) = \log_2\left(\frac{C(X \to Y)N}{C(X)C(Y)}\right)$$

where $C(X)$ and $C(Y)$ are the number of articles of X and Y, and $C(X \to Y)$ is the number of articles of X that cite articles of Y. Note that $PMI(X \to Y)$ is not necessarily equal to $PMI(Y \to X)$.

It is obvious that the PMI values between any two of these 3120 subject types can be directly obtained by simple analyses of the WoS via the above formula, but there are two major limitations that hinder the efficiency and the feasibility of this calculation. First, when a database or a network is too large for a personal computer to analyse, the calculation may spend unexpectedly long computational time to finish the calculation. It is exactly the case when this calculation is applied to the large-scaled citation database like the WoS. The computational time is too long for one to request a daily update on all PMI values. Second, when a subject type is newly introduced to the WoS in the future, its PMI value with any other subject types will be significantly underestimated due to the lack of adequate amount of counts. Similar situation may happen to some small subject types that are suspicious to be recently introduced to the WoS.

Classification via Deep Learning and the Setup for Subject Type Dependency

It is of great interest to consider an efficient and accurate machine to estimate the PMI value between two subject types in the WoS, or to predict it when at least one subject type is newly introduced or even does not exist in the WoS. In this section, we introduce a classification method to determine if two subject types are independent via deep learning. If they are determined as independent, then their PMI value is $-\infty$ (usually denoted as NA). If not, we consider to estimate their PMI value via a regression model or any other appropriate models.

A Brief Review on the Classification Methods in the Literature

There are many classification methods in the literature, but none perform well in all problems. Thus, conventional wisdom suggests to choose the appropriate method based on the nature and property of the data. Here are some popular supervised learning algorithms.

K-nearest neighbour (k-NN) algorithm (Altman 1992) is a non-parametric method in pattern recognition that classifies an instance via a majority vote among its k nearest neighbours.

It is a type of instance-based learning where the function is only approximated locally and all computation is deferred until classification. However, it is computationally expensive in the testing phrase that is impractical in many real applications. It is also hard to implement in high-dimensional feature space, so this algorithm is not suitable for our purpose.

Naïve Bayes (NB) classifier (Kononenko 1993) is a simple probabilistic classifier with strong conditional independence assumptions between all features. In specific, it takes prior probability into account and assigns the instance to the class with maximum posterior probability based on Bayes' rule. Although it is fast and easy to implement, it lacks of an ability to learn interactions between features.

Decision tree (Breiman et al. 1984) is a decision support tool that recasts all decisions and their possible consequences as a hierarchical structure. It is a popular tool in machine learning and decision analysis to identify a goal-oriented strategy. Tree is split into branches based on features that returns the highest information gain. This method is preferred due to its interpretability, but it is often difficult to generalize. To improve from the obstacles in the original framework, ensemble methods like random forest appears. Linear discriminant analysis (LDA) aims at finding the linear combination (class) of features that maximizes the variation between classes while minimize the variation within classes. However, its assumptions on Gaussian-distributed data and equal variance on attributes may not be valid in every practical scenario. Support vector machines (SVM) aims at finding a hyperplane with maximal margin that separates different classes of features. Although it can be extended to nonlinear scenario easily using kernel function, it may become slow for large-scale data because it is a batching learning method. Further details on these classifiers are referred to Dixon and Brereton (2009).

The General Framework of Deep Learning on Classification Problems

Deep learning have received much attentions among computer scientists for the past ten years. It is a broader family of machine learning methods on learning data representations, so it has a wide variety of applications in many different fields including image processing, signal recognition, natural language processing, bioinformatics, and many others. See Deng and Yu (2014) for detailed introductions. There are many different architectures in deep learning, like multi-layer perceptron (MLP), convolutional neural net (CNN) and Recursive neural network (RNN). Different architectures are used to solve different types of problems.

In this work, we adapt the multilayer perceptron (MLP), which is a particular type of artificial neural network to handle nonlinearly separable data. MLP is composed of three consecutive layers in the order of input, hidden and output layers, and each layer consists of some nodes, which are also known as neurons. Since all nodes in MLP are expected to be fully connected, each node in a layer is connected with a linear combination of the nonlinear functional transformation of all nodes of its previous layer. The nonlinear function is generally called the activation function, and some common choices include the sigmoid function, rectified linear unit (ReLU) function, or inverse trigonometric functions. Based on back-propagation, MLP adjusts the weights and bias repeatedly such that the loss function can be minimized. In general, it tends to perform better when a problem with high dimensional and continuous features is handled.

Input and Output Layers

Given two subject types X and Y, which each can be specified in the form of a vector of length 275:

$$X = (x_1, x_2, \cdots, x_{275}), Y = (y_1, y_2, \cdots, y_{275})$$

where the entries x_i and y_i are Boolean indicators whether subject i is a component listed in X and Y respectively. Since the response variable $PMI(X \to Y)$ indicates a specific direction from X to Y, we create a change variable

$$\Delta XY = \left(\delta(x_1 \to y_1), \delta(x_2 \to y_2), \cdots, \delta(x_{275} \to y_{275})\right)$$

where $\delta(x_i \to y_i)$ is a code on the change of inclusion of subject i from subject type X to subject type Y, and the codes are mapped below:

$$\delta: (0 \to 0) \to 0;\ (0 \to 1) \to -1;\ (1 \to 0) \to 1;\ (1 \to 1) \to 2.$$

The first three map is trivial and we indicate 2 instead of 0 in the last map so that we can distinguish two different "unchanged" states.

We then rephrase from the problem of determining the subject type dependency as follows. The goal of this classification is to determine if two subject types are independent or not. Therefore, we begin the input layer with 550 features, which include the first 275 features that are the indicators of subject components of X, and the last 275 features that are the codes of subject component change ΔXY. The output layer is certainly the Boolean indicator of the response $PMI(X \to Y)$, where 0 stands for independency and 1 stands for the opposite.

Other Setups in MLP

We conduct a MLP with batch size of 10000 and 5 epochs. We use the cross entropy as the loss function and the adaptive moment estimation as the optimizer. There is no theoretical justification on the number of layers and number of hidden neurons in each layer being used in a MLP. A standard practice from common software suggests four layers. In the next section, we follow the standard practice and use four layers in the first trial, and we reduce to use three layers in our final trial with solid explanations on the layer reduction.

There are two activation functions in the MLP. We use the ReLU function as the major activation function in between layers and its functional form is

$$f(x_i) = max(0, x_i)$$

Due to the probabilistic nature of the indicator in the output layer, we use the softmax function as the activation function and its functional form is

$$f(x_i) = e^{x_i} / \sum_k e^{x_k}.$$

Results on the Study of Subject Type Dependency

Results of 4-layer MLP

As discussed above, the number of input features is 550 and the number of output indicators is 2. The 4-layer MLP is first considered in the first trial. The scheme is illustrated in Figure 1.

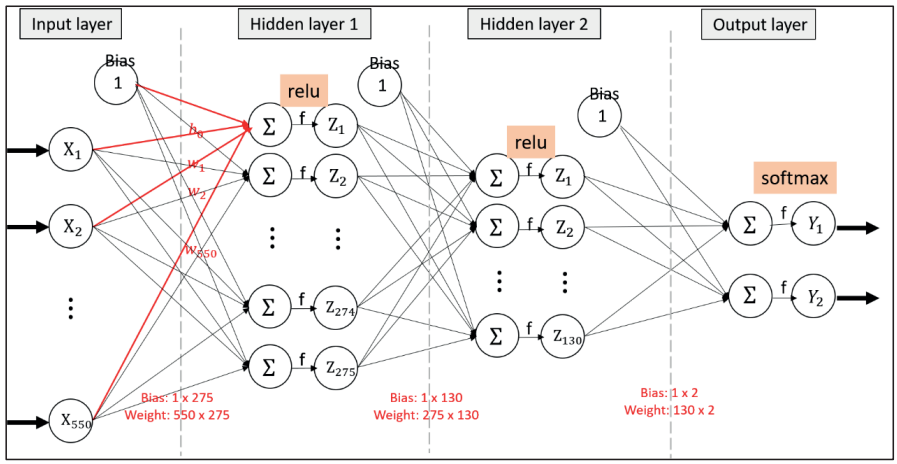


Figure 1. The scheme of 4-layer MLP

We start with obtaining each of the 275 nodes in the first hidden layer from the linear combinations of all 550 nonlinearly transformed features in the input layer via the ReLU activation function. Then we obtain each of the 130 nodes in the second hidden layer from the linear combinations of all 275 nonlinearly transformed nodes in the first hidden layer via the ReLU activation function. Finally, we obtain our classification results in the output layer from linear combination of all 130 nonlinearly transformed nodes in the second hidden layer via the softmax function instead of ReLU function, so that we can interpret the outcomes as a probability. The classification result is shown as a confusion table in Table 1.

Table 1. The Confusion Table on the Classification Results of 4-Layer MLP.

		True Dependency Status	
Predicted Dependency Status		0	1
	0	81.98% (TP)	6.18% (FP)
	1	4.06% (FN)	7.78% (TN)

There are a total of 9731280 pairs from 3120 subject types being investigated, and 86.04% of these pairs are truly independent. The true positive rate is 95.28%, meaning that 95.28% of the true independent pairs are successfully detected by the 4-layer MLP. The precision is 92.99%, meaning that when two subject types are classified as independent by the 4-layer MLP, 92.99% is accurate detection. The F_1 score, which is the harmonic mean of precision and sensitivity, is 0.9412.

Notice that the goal of this deep learning step is to determine if two subject types are independent or not, so we focus only on the results of zeros predicted in the output layers. When the output indicator suggests that two subject types are not independent, they are moved into the next step for the estimation or prediction of PMI values.

The illustration of all weights of three combinations in between four layers is too complex and high-dimensional. Figure 2 shows the weights of randomly chosen 5 nodes on the first two linear combinations.

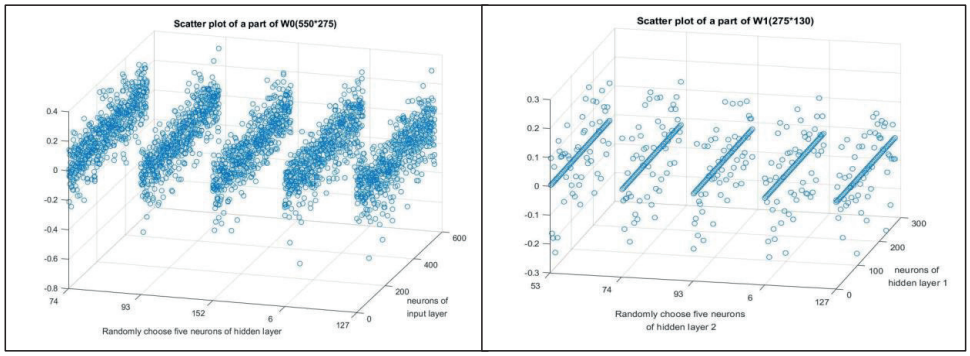


Figure 2. (left) Weights between the input layer and the 1st hidden layer in 4-layer MLP; (right) Weights between two hidden layers in 4-layer MLP.

Figure 2(right) shows some line-type patterns in the weights of every node, and these patterns correspond to a large amount of zero weights in the combinations. We believe that it is an evidence of the usage of too many hidden layers in MLP, which is a similar concept of overfitting in the regression modelling. Thus, a 3-layer MLP is used in the final trial.

Results of 3-layer MLP

We use the same setting as the 4-layer MLP except we only consider one hidden layer with 275 nodes. The scheme is illustrated in Figure 3.

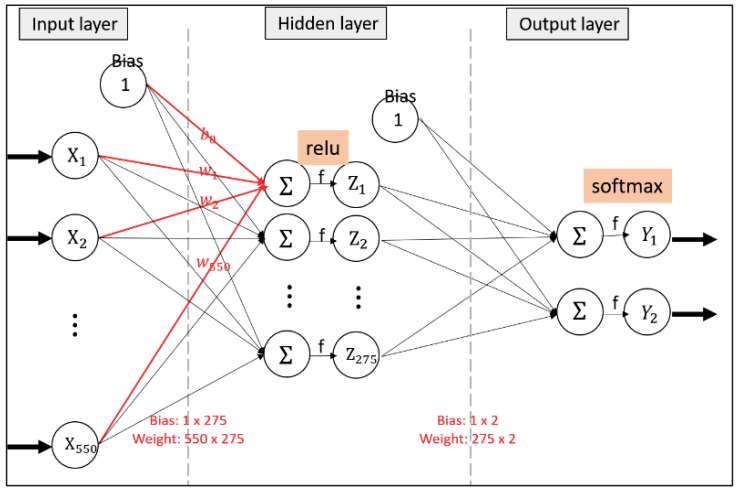


Figure 3. The scheme of 3-layer MLP

We start similarly to obtain each of the 275 nodes in the hidden layer from the linear combinations of all nonlinearly transformed 550 features in the input layer via the ReLU activation function. Instead of having the second hidden layer, we obtain our classification results in the output layer directly from the linear combination of all 275 nonlinearly transformed nodes in the hidden layer via the softmax activation function. The classification result is shown as the confusion table in Table 2.

Table 2. The Confusion Table on the Classification Results of 3-Layer MLP.

		True Dependency Status	
Predicted Dependency Status		0	1
	0	81.96% (TP)	6.28% (FP)
	1	4.08% (FN)	7.68% (TN)

The performance of 3-layer MLP is as good as that of 4-layer MLP from the confusion table above. The true positive rate is 95.26% and the precision is 92.88%, which are just slight

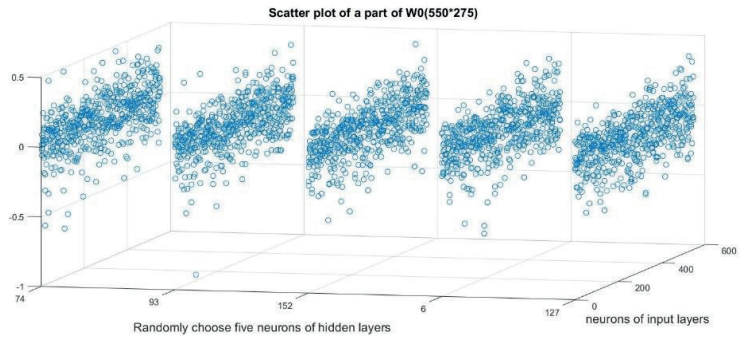


Figure 4. Weights between the input layer and the hidden layer in 3-layer MLP.

decreases of 0.02% and 0.11% respectively from a MLP with an additional hidden layer. The F_1 score is 0.9406, a 0.06% decrease from the F_1 score of 4-layer MLP. Considering the 3-layer MLP is structurally less complex than its 4-layer counterpart, we believe it is appropriate to use the 3-layer MLP as the efficient deep learning classifier to determine whether two subject types are independent.

The complete weighting matrix from the input layer to the hidden layer is of dimension 550×275, and we randomly select five nodes in the hidden layer and illustrate the weights of all 550 features towards these five nodes in Figure 4. In addition, the weighting matrix from the hidden layer to the output layer is of dimension 275×2 and we show how the weights of these 275 nodes are combined to obtain the eventual output indicators in Figure 5.

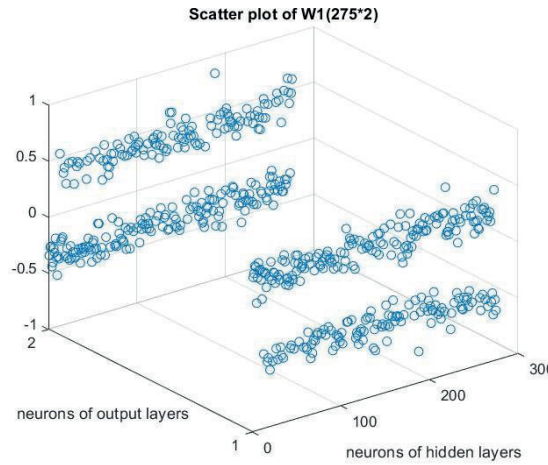


Figure 5. Weights between the hidden layer and the output layer in 3-layer MLP

A Demonstration

We try to quantitative evaluate the statement posted in the beginning of this work: "Applied Mathematics and Physics should be very close to one another and both subjects have little relationship with Literature".

We first look into the WoS database and find that ID2637 stands for "Mathematics + Applied Mathematics", which is the closest subject type to Applied Mathematics. In addition, ID2918 and ID2531 stand for "Physics" and "Literature" respectively. Table 3 shows the actual PMI values obtained from the WoS database and Table 4 shows the results on the determination of subject type dependency using the 3-layer MLP. If the determination is perfect, any numerical values in Table 3 will match the indicator of "Not Independent" in Table 4, while the "NA" state in Table 3, which implies no citation count is found from one subject to another, will match the indicator of "Independent".

Table 3. Actual PMI Values among Three Subjects from the WoS Database.

			Y	
	$PMI(X \rightarrow Y)$	Applied Math.	Physics	Literature
	Applied Math.		0.1893	-2.2029
X	Physics	0.8122		NA
	Literature	-5.0212	NA	

Table 4. Predicted Dependency Status among Three Subjects from the WoS Database.

		Y		
	$PMI(X \rightarrow Y)$	Applied Math.	Physics	Literature
X	Applied Math.		Not Independent	Not Independent
	Physics	Not Independent		Not Independent
	Literature	Independent	Independent	

The above two tables show the correct determination on the following four pairs:

1. "Applied Mathematics → Physics" (0.1893 and "Not Independent");
2. "Applied Mathematics → Literature" (-2.2029 and "Not Independent");
3. "Physics → Applied Mathematics" (0.8122 and "Not Independent"); and
4. "Literature → Physics" (NA and "Independent).

For the pair "Literature → Applied Mathematics", although the actual PMI value is numerical, it is so negatively large that the corresponding citation count is very small. Thus, this subject pair can be viewed as barely dependent, and the false determination is possible. For another pair "Physics → Literature", the 3-layer MLP fails to determine their independency, but this pair will enter the modelling part together with the remaining 13.96% of total number of pairs. If the regression modelling is appropriate, this pair is still possible to be estimated as independent.

Discussion and Conclusion

In this work, we propose a deep learning approach to determine if two subject types are independent or not based on the subject components of two subject types. Different from the naïve approach to calculate PMI values via citation counts obtained from the WoS database, our proposed approach is efficient and provides the classification when one of the subject types or both are do not exist in the database.

This deep learning approach is built on the architecture of multilayer perceptron (MLP). One may question on several common arguments on the use and the setting of MLP. For example, although we justify why the 3-layer MLP is suggested rather than the 4-layer MLP in our work, there is a lack of valid test or criterion to justify the best number of layers in general. Similar question also appears in the choice of the number of nodes in the hidden layers. In addition, one may also question on the choice of the activation function and many other settings. Nevertheless, the main goal of this work is to demonstrate how deep learning can help to determine the subject type dependency efficiently. Further studies on the technical details of the MLP is out of the scope of this work.

The approach can be considered as the first step to the estimation and prediction of the PMI values between two subject types. Notice that this work only tells whether two subject types are independent or not. If they are independent, we can safely say that the PMI value is zero. However, if this approach suggests the dependency, we will need a follow-up analysis on the estimation of the PMI value between two subject types that exist in the WoS, or the prediction of the PMI value between two subject types that at least one does not exist in the WoS. Therefore, another implicit contribution of this work is to suggest an efficient method to greatly reduce the data size to about 11.76% of the complete data, which are the pairs that the 3-layer MLP determines as dependent in Table 2.

In the next step, a simple way to conduct the analysis on the PMI value from X to Y is to perform a simple linear regression analysis. We consider to use 550 features, which includes 275 subject components of X and 275 change components from X to Y, as independent factors and the response is the PMI value. Figure 6 shows the magnitudes of all 550 regression coefficients.

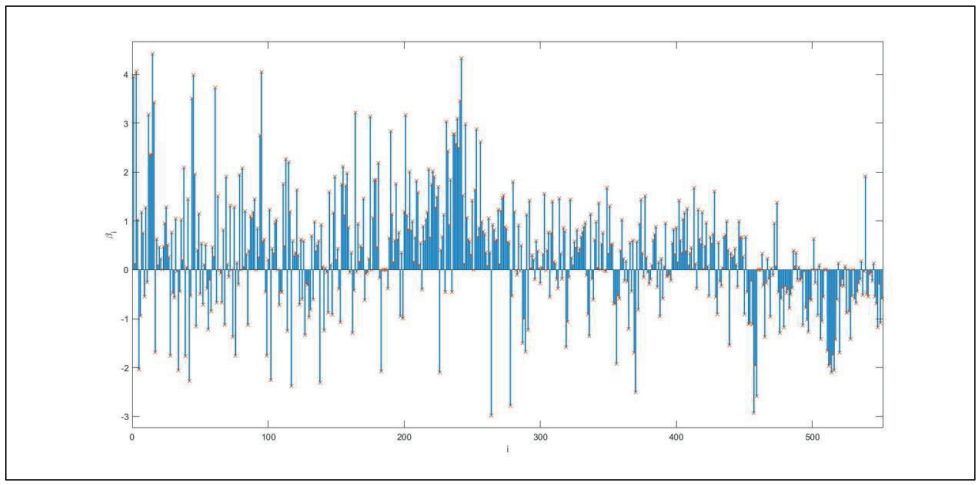


Figure 6. The magnitudes of all 550 regression coefficients in the regression model.

In practice, we can quickly compute the PMI values by inputting the subject components of two subject types. Table 5 shows some comparisons between the actual and estimated PMI values.

Table 5. Comparison between Actual and Estimated PMI Values.

	Actual PMI Values	Estimated PMI Values
Applied Math. → Statistics	6.6200	6.0422
Statistics → Applied Math.	6.6537	5.7104
Applied Math. → Physics	0.1893	3.0027
Physics → Applied Math.	0.8122	2.9688

It is obvious that the estimation of the PMI values is not always good. There are many cases like the PMI values between Applied Mathematics (ID2637) and Statistics (ID2670) that their PMI values are estimated successfully. There are also some cases like the PMI values between Applied Mathematics and Physics (ID2918) that the estimated values are seriously deviated.

There are several reasons to explain this failure. First, Table 2 states that among all pairs of subject types that the 3-layer MLP determines as dependent, 34.69% of them are actually independent. It implies that there are excessive amount of zero counts on the citations, and it violates the fundamental assumption of normal errors in the simple linear regression. One possible solution is to consider the zero-inflated Poisson (ZIP) regression, but this leads to a whole new research on the implementation and modelling under this new framework. We consider this as the future work from this work.

Acknowledgments

The authors would like to thank Clarivate Analytics to provide the access to the raw data of the Web of Science database for research investigations. They also thank the URA team of ISM for transforming the data into neo4j database and providing the neo4j database for analysis in this work. This work was supported by the Ministry of Science and Technology (Taiwan) grant numbers 107-2118-M-001-011-MY3 and 107-2321-B-001-038.

References

Altman, N.S. (1992). An introduction to kernel and nearest-neighbor nonparametric regression. *The American Statistician*, 46, 175-185.

Ascoli, G.A. (Ed). (2002). *Computational Neuroanatomy: Principles and Methods*. Totowa, New Jersey: Humana Press.

Bakshi, U.A. & Bakshi, A.V. (2008). *Electrical Networks*. Pume, India: Technical Publications.

Breiman, L., Friedman, J., Stone, C.J. & Olshen, R.A. (1984). *Classification and Regression Trees*. Belmont: Wadsworth.

Chang, L.L.N., Phoa. F.K.H. & Nakano, J. (2018+). A new metrics for the analysis of the scientific article citation network. *In Review*.

Church, K.W. and Hanks, P. (1990). Word association norms, mutual information and lexicography. *Computational Linguistics*, 16, 22-29.

Deng, L. & Yu, D. (2014). Deep learning: methods and applications. *Foundations and Trends in Signal Processing*, 7, 1-199.

Dixon, S.J. & Brereton, R.G. (2009). Comparison of performance of five common classifiers represented as boundary methods: Euclidean distance to centroids, linear discriminant analysis, quadratic discriminant analysis, learning vector quantization and support vector machines, as dependent on data structure. *Chemometrics and Intelligent Laboratory Systems*, 95, 1-17.

Kononenko, I. (1993) Inductive and Bayesian learning in medical diagnosis. *Applied Artificial Intelligence*, 7, 331-337.

Phoa, F.K.H. & Sanchez, J. (2013). Modeling the browsing behaviour of world wide web users. *Open Journal of Statistics*, 3, 145-154.

Roder, M., Both, A. & Hinneburg, A. (2015). Exploring the space of topic coherence measures. *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining* (ACM WSDM '15) (pp. 399-408). New York: ACM.

Wang, T.C. & Phoa, F.K.H. (2016). A scanning method for detecting clustering pattern of both attribute and structure in social networks. *Physica A*, 445, 295-309.

Yu, Y., Dong, J., Liu, K.J. & Nakhleh, L. (2014). Maximum likelihood inference of reticulate evolutionary histories. *Proceedings of National Academy of Sciences of the USA*, 11146:16448-16453.

Persistence of journal hierarchy in open access publishing

Vincent Traag and Ludo Waltman
Centre for Science and Technology Studies (CWTS), Leiden University, the Netherlands

Abstract

The recently proposed *Plan S* envisions a shift to an academic publishing system that is fully open access. The plan has led to heated debates on many aspects of academic publishing. There are for example concerns that high-quality journals will no longer be viable. The argument is typically that journals can increase their profits by lowering their quality standards. We here investigate this argument by analysing a simple model of the publishing system. We find that journals continue to have an incentive to maintain a certain quality in order to attract more submissions. Hence, a distinctive journal hierarchy persists, even if article processing charges (APCs) are capped to a maximum. Nevertheless, a cap on APCs may have significant consequences. The lower the cap, the smaller the quality differences between journals.

Introduction

Open access has been on the rise for quite some years. Recently, eleven European research funding agencies proposed *Plan S*[1], in which they announce that all research they fund should be published in open access journals, and more funding agencies have joined since. Plan S does not allow publishing in hybrid journals, which combine an open access model and a subscription model, except for a transitionary period. This raised concerns about the continuing existence of high-quality journals, and academic society-owned journals in particular. For example, the Global Young Academy (2018) argued that in a negative scenario "there is a strong profit incentive for publishers to favour quantity over quality". Similarly, a number of scholars claimed that Plan S "stimulate[s] accepting as many papers as possible—regardless of their quality" (Research Community, 2018). Other academic societies argued that Plan S "actually incentivizes publishers to go after more and more papers" (Brainard, 2019).

We here focus on the possibility of dwindling qualities of journals as a result of Plan S. Many other issues from Plan S are being debated, such as the issue of copyright, the effect on scientific careers for young scholars, the access to publishing for poorer nations and the claimed restrictions on academic freedom. We do not address these issues here, although they should of course play a role in the discussion on Plan S.

We propose a relatively simple model of open access publishing, which is sufficiently realistic to be useful. We restrict ourselves to open access publishing in which authors need to make a payment to publish their article, commonly called an article processing charge (APC). This form of open access publishing is often referred to as gold open access. Other options, such as self-archiving, sometimes referred to as green open access, or open access journals without APCs, sometimes referred to as platina open access, will have different dynamics. We believe our model is useful in sharpening the ideas about possible consequences and effects of Plan S. We first introduce the model, then show some dynamics and outcomes of the model, and finally discuss the implications for Plan S.

Model

Let there be n different journals 1, …, n, each with an associated quality threshold q_j. We assume there are N articles, and the quality of each article is drawn from some distribution. Authors benefit from publishing their articles in high-quality journals, and they therefore try to publish in these journals. However, different authors have different perceptions of the quality of journals. For each article, the author of the article assesses the quality of each of the

[1] https://www.coalition-s.org/

n journals in which he or she could try to publish the article. An author's perceived quality u_j of journal j equals

$$u_j = \beta q_j + \epsilon_j,$$

where all ϵ_j are assumed to be independently distributed according to the Gumbel distribution, for which $\Pr(\epsilon_j > s) = e^{-e^{-s}}$. An author will first submit his or her article to the journal j with the highest perceived quality u_j. The probability that journal j has the highest perceived quality u_j is a standard result in discrete choice theory (Anderson, Palma, & Thisse, 1992). This probability equals

$$\frac{e^{\beta q_j}}{\sum_{i=1}^{n} e^{\beta q_i}}.$$

If an article is rejected by this journal, it will be submitted to the journal k with the next-highest perceived quality u_k, and so on. We can again work out the probability that an article will be submitted to journal k after it has been rejected by some other journal j. For $\beta = 0$, journal quality effectively does not play a role in the order in which authors submit their articles to different journals. For $\beta \to \infty$, authors always submit in the order of decreasing journal quality.

When an article is submitted to journal j, the journal accepts the article if the article quality is above the threshold q_j. Generally, the probability that an article has a quality higher than q_j is $\Pr(Q > q_j)$, where Q is the random variable denoting the quality of an article. We assume that $Q \sim \text{LogNormal}(\frac{-1}{2}, 1)$, so that the average Q is 1.

In a system of gold APC-based open access publishing, the expected total revenue of a journal j is simply the expected number of accepted articles $E(A_j)$ times the APC charged by the journal. Following suggestions made in Plan S, we assume that a cap is imposed on APCs. Since APCs are paid by research funders, not by individual authors, each journal will have an incentive to set its APC equal to this cap. We therefore assume that each journal has the same APC. This APC equals the cap. The total expected revenue of a journal is then equal to $\text{APC} \cdot E(A_j)$. We assume that the cost of running a journal consists of some fixed costs c_f, some variable cost c_s that scale with the number of submissions, and variables costs c_a that scale with the number of accepted articles. These costs are all independent of the quality threshold q_j. The expected profit is then

$$E(X_j) = \text{APC} \cdot E(A_j) - c_f - c_s \cdot E(N_j) - c_a \cdot E(n_a)$$

It is possible to calculate for each journal j the expected number of submitted articles $E(N_j)$ and the expected number of accepted articles $E(A_j)$ (i.e. published articles). We do not include these calculations here for reasons of brevity.

Quality evolution

We assume that journals are focused on maximizing their profit. Each journal adjusts its quality threshold q_j in order to increase its expected profit. We consider it unrealistic to assume that the quality threshold can be set to an arbitrary level. There is a high degree of inertia in the journal system, and authors are likely to respond to realized quality, rather than

the quality threshold q_j. In order to model this, we assume that a journal is only able to slightly adjust its quality threshold. In particular, we assume that a journal adjusts its quality threshold q_j according to the gradient of the profit

$$\frac{\partial E(X_j)}{\partial q_j}.$$

We again do not include the calculations here for reasons of brevity.
One may object that journals are not purely driven by profit. Some journals may have an intrinsic drive to uphold a certain quality, simply to contribute to high-quality science. Of course, such an intrinsic drive will be easier to uphold when it is aligned with pressure exerted by considerations of profitability. Hence, even though we acknowledge the existence of such an intrinsic desire for quality, we believe that for many journals it will be difficult to sustain a certain quality on the sole basis of this motivation in the setting of gold APC-based open access. Academic publishing is a market with high profit margins and clear commercial interests, which are unlikely to play only a minor role. Moreover, exactly because of the high degree of inertia mentioned above, and the vested interests many researchers have in publishing in leading journals that are often owned by commercial parties, it is likely that commercial considerations will continue to exert considerable pressure on journals.

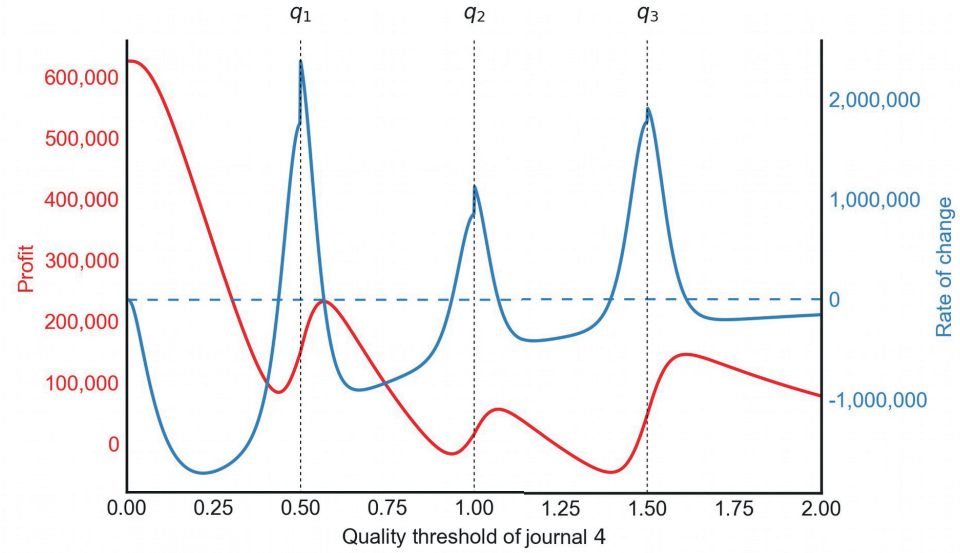


Figure 1. Profit versus quality threshold. The figure shows the profit of journal 4 for a certain quality threshold (in red), given the quality thresholds of three other journals. It also shows the rate of change of the profit for a certain quality threshold (in blue). We set $\mathrm{APC}=2000$**,** $c_f=10000$**,** $c_s=100$**,** $c_a=400$**,** $\beta=30$ **and** $N=1000$**.**

Results

We first illustrate the model by plotting for a specific journal how the profit and the rate of change of the profit depend on the journal's quality threshold (Figure 1). We assume there are four journals, of which the first three have quality thresholds of $q_1=0.5$, $q_2=1$ and $q_3=1.5$. Figure 1 shows (in red) the profit of the fourth journal as a function of its quality threshold q_4. Overall, the profit is highest for q_4 near 0. However, if the journal can adjust its quality

threshold only slightly, it will adjust its quality threshold depending on the current value of the threshold. For example, assume that the quality threshold q_4 is slightly less than 1. The journal then has an incentive to increase its quality threshold so that it becomes slightly larger than 1, thereby attracting more submissions of sufficiently high quality. These submissions would have otherwise gone to journal 2, which has a quality threshold of $q_2=1$. When q_4 is slightly higher than 1, journal 4 no longer has an incentive to change its quality threshold. If the journal decreases its quality threshold, it will attract fewer submissions, thereby decreasing its profit. If it increases its quality threshold, it will not attract substantially more submissions, but it will accept fewer articles, thereby decreasing its profit. This can also be seen by the rate of change in Figure 1 (in blue). When the rate of change is positive, journal 4 can increase its profit by increasing its quality threshold, and when the rate of change is negative, the journal can increase its profit by decreasing its quality threshold. This illustrates how journal 4 adjusts its quality threshold. However, the other three journals will make similar adjustments, and the overall dynamics are therefore more complex.

With two journals that both adjust their quality thresholds, the dynamics are already quite complex. In Figure 2, we show the dynamics from three different perspectives. In Figure 2A, we show for each combination of q_1 and q_2 the local direction in which the two journals adjust their quality thresholds to try to increase their profits. Note that the dynamics are entirely symmetrical, and the upper left part of the plot mirrors the lower right part. We show how the dynamics play out from one particular starting point $q_1=0.4$ and $q_2=0.3$. In Figure 2B, we show the temporal evolution of the quality thresholds starting from this particular starting point $q_1=0.4$ and $q_2=0.3$. In Figure 2C, we show the temporal evolution of the expected profits. Journal 1 initially increases its quality threshold, while journal 2 immediately decreases its quality threshold. At some point, journal 1 also starts to decrease its quality threshold, while at some later point journal 2 starts to increase its quality threshold again. Although one may expect that the profits always increase, this is not the case. At first the profits increase for both journals, but after some time they decrease again, which is most clearly visible for journal 1. This happens because journal 2 increases its quality threshold, resulting in a lower profit for journal 1.

The dynamics for more journals become more complex, as shown in Figure 3. Even though differences in the quality thresholds of journals may decrease, there is a persistent hierarchy of journals in terms of quality standards. The ranking of journals relative to each other is preserved over time. Figure 3A shows that the highest quality journal converges to a distinctively higher quality threshold than the second-highest quality journal. Journals clearly do not have an incentive to continue decreasing their quality thresholds in order to increase their profits. Finally, it may be of some interest to note that the highest-quality journals have the highest profits, as shown in Figure 3B. The highest quality journals are also the ones with the highest rejection rates (between 80-90%), which seems quite realistic. Perhaps somewhat unrealistically, high-quality journals publish most articles, and many of the low-quality journals publish only a small number of articles.

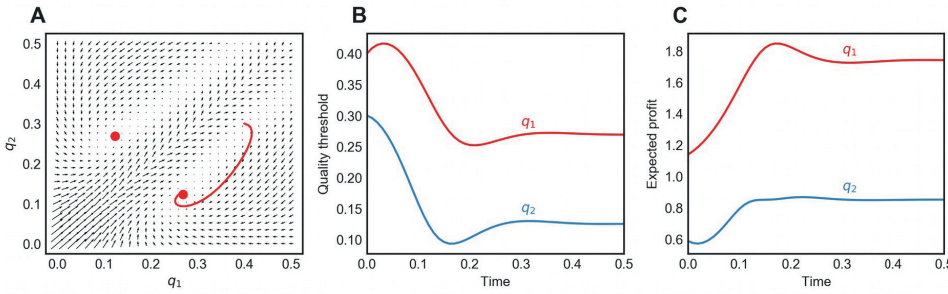


Figure 2. Dynamics of two journals that adjust their quality thresholds in order to try to increase their profits. We set APC$=5$, $c_f=0$, $c_s=1$, $c_a=1$, $\beta=10$ **and** $N=1$.

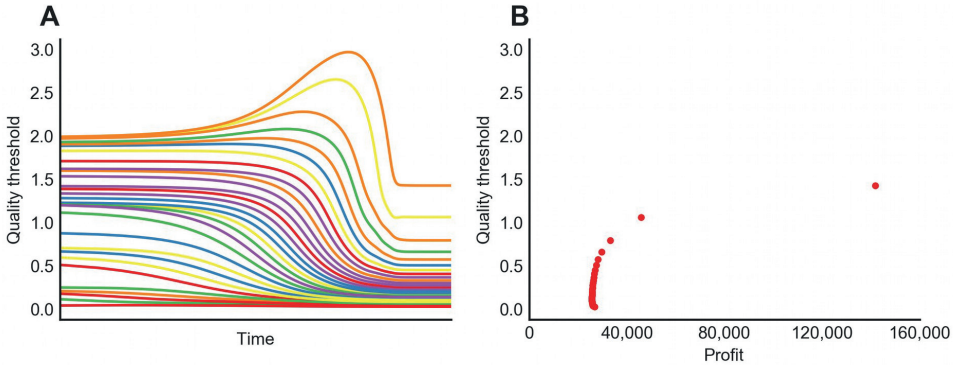


Figure 3. Dynamics for 30 journals for some random initial condition, and the profits of all journals after the dynamics have converged. We set APC $=2000$, $c_f=10000$, $c_s=100$, $c_a=400$, $\beta=30$ **and** $N=1000$.

Finally, in Figure 4, we analyse the effect of the APC level on the quality thresholds to which four journals converge. If the APC is insufficiently high, all four journals converge to a quality threshold of zero. At this point, the APC is lower than the cost of receiving submissions and publishing articles, so that all journals make a loss. When the APC is sufficiently high, a journal hierarchy emerges. Higher APCs result in an increasingly distinctive journal hierarchy.

Discussion

The debate on open access has received an impulse by the recently proposed Plan S. One particular concern is whether high-quality journals may continue to exist in the publishing system envisioned in Plan S.

We have addressed this question using a theoretical model in which journals are assumed to gradually adjust their quality standards in order to increase their profits. We find that high-quality journals may continue to exist, provided that a cap imposed on APCs is sufficiently high. Possibly, the quality standards of journals may go down. However, journals will maintain certain quality standards in order to continue attracting submissions. This effect is the result of authors preferring to publish in high-quality journals, and preferably submitting their work there. The typical argument that journals will increase profits by simply accepting more articles, and hence lowering their quality standards, does not consider this effect. Given

the theoretical nature of our model, it is difficult to obtain a concrete estimate of the cap on APCs that would be needed.

An issue that our model does not take into account is that many authors who are not funded by research funding agencies that support Plan S will consider the level of the APC of a journal when deciding where to submit their articles (West, Bergstrom, & Bergstrom, 2014). This may cause APCs to become more differentiated, with higher quality journals charging higher APCs than lower quality journals. If authors have no incentive to consider the level of the APC of a journal, for example because the APC will be paid by research funders that support Plan S, journals will simply raise their APCs until they hit the cap imposed by research funders. This creates a dilemma for research funders. If it is considered important to have high-quality journals, funders should set a relatively high cap on APCs. However, low-quality journals will then also benefit from high APCs, leading to a costly publishing system.

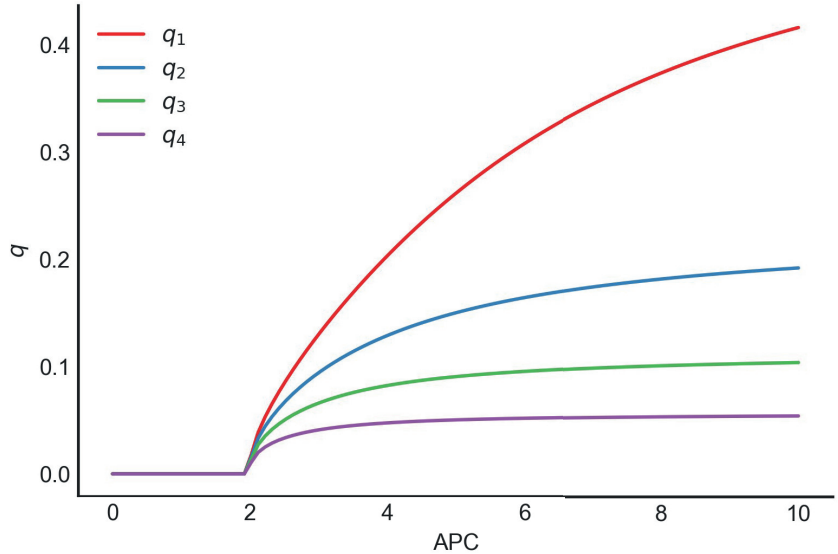


Figure 4. Journal hierarchies resulting for different APCs. We set $c_f=0$, $c_s=1$, $c_a=1$, $\beta=3$ and $N=1$.

To reduce costs, funders could set the cap on APCs at a relatively low level. Although a journal hierarchy could then still persist, quality differences between journals will decrease and journals that have very demanding quality standards are unlikely to uphold them.

We would like to suggest two ways in which funders could deal with the above dilemma. First, funders could differentiate the cap on APCs based on the services provided by journals. This requires journals to be transparent about the services they offer and the associated costs. In the current Plan S implementation guidance, such transparency is indeed already required. Second, rather than covering APCs for all articles, funders could include a budget for APCs within the overall funding they provide to researchers. Researchers then have to weigh the APC against the quality of a journal when deciding where to submit their manuscript. This would make researchers more sensitive to the level of the APC of a journal, leading to more differentiated APCs.

References

Anderson, S. P., Palma, A. de de, & Thisse, J.-F. (1992). *Discrete Choice Theory of Product Differentiation*. Cambridge: The MIT Press.

Brainard, J. (2019). Scientific societies worry Plan S will make them shutter journals, slash services. *Science*. https://doi.org/10.1126/science.aaw7718

Global Young Academy. (2018). *Opportunities and challenges for implementing Plan S*. Retrieved from https://globalyoungacademy.net/wp-content/uploads/2018/10/YA-Statement-on-Plan-S-FINAL.pdf

Research Community. (2018). Reaction of Researchers to Plan S. https://doi.org/10.5281/ZENODO.1477914

West, J. D., Bergstrom, T., & Bergstrom, C. T. (2014). Cost effectiveness of Open Access publications. *Economic Inquiry*, *52*(4), 1315–1321. https://doi.org/10.1111/ecin.12117

Analysis of Division of Labor in High Quality Life Science Research of China

Tao Han[1] and Xiaoyu Cai[2]

[1]*hant@mail.las.ac.cn* [2]*caixiaoyu@mail.las.ac.cn*

National Science Library, Chinese Academy of Sciences, 33 Beisihuan Xilu, Zhongguancun, Beijing 100190 (P. R. China)

Department of Library, Information and Archives Management, School of Economics and Management, University of Chinese Academy of Sciences, No.19(A) Yuquan Road, Shijingshan District, Beijing 100049 (P. R. China)

Abstract

International cooperation has always been a hotspot in academic research. Recently, with the rise of 'Author Contributions', quantitative study using authors' division of labor in scientific research provide a new perspective for the measure of international cooperation.. In this article, semantic structures of division of labor are extracted with text mining from the author contribution part of life science publications in Nature Index 68, and identified with two predefined rule libraries into the corresponding division category. Then quantitative analysis of Chinese division of labor is conducted from both quantity and structure perspectives, at both national and international level.From quantity perspective at both national and international level, Chinese contributions in any category of division of labor has been increasing in recent 10 years not only in the proportion but also in the rank among the world, especially Experiment Operation. From the perspective of domestic structure, the focus of division of labor has shifted to Data Analysis and Management in the past decade. From the comparative analysis of international structure, except for South Africa, China has similar division of labor with other 10 countries. From the co-authorship perspective, the co-authorship model of China, USA and Canada is relatively similar, while the difference between co-authorship models of China and other countries are mainly reflected on Sample and Data Collection and Tools and Technology.

Introduction

Nature Index, proposed in 2014, aims to focus on a relatively small number of high-quality papers. Although the 68 journals of Nature Index selected are less than 1% of the journals cited in the Journal Citation Reports, the proportion of their papers' citations is close to 30%. Therefore, publications in Nature Index can effectively stand for high-quality research output. Author contribution is a short text usually behind the main body of a literature, which describe the category of work each author undertakes. Author contribution is also called authorship or contributorship. The discussion of author contribution can be traced back to 1999. V. Yank and D. Rennie (1999) designed a small-scale experiment in which they proposed that it is feasible to use the "Author Contributions" to analyse the author's real contribution in the paper. And it deserved to be attached importance to by editorial departments and scholars. In 1999, the reviewer of Nature proposed to mark 'Author Contributions' in the paper, which means that journals put 'Author Contributions' into practice. Since then, more journals have gradually requested the 'Author Contributions' in journal publications. Up to 2017, 44 of the 68 journals in Nature Index have the author contribution structure (we argue that a journal has author contribution if the amount of paper which has 'Author Contributions' exceeds 10% in the whole issue). The emergence of 'Author Contributions' not only provides a new perspective to understand the author's value, but also provides a new perspective to analyse the division of labor in scientific cooperation. Previous studies which based on the 'Author Contributions' part of journals mostly focus on the quantitative calculation of the author's personal contribution (Feeser et al. 2014, Zhang. 2016), or on the distribution characteristics of authors' division of labor (Larivière et al, 2016). However, less research has analyzed the author contribution from the national level, to be more specific, measured the international cooperation using the authors' division of labor in author contribution. This study will contribute to filling this gap. The emergence of international cooperation has promoted the development of scientific research. The international collaboration has been growing with a peculiar pattern-faster than an exponential growth, shaping a scale-free network (Ribeiro L C et al, 2017). International

collaboration can not only enhance the forefront of research questions, broaden the research ideas and improve the research capacity of researchers, but also promote the development of vulnerable discipline and traditional preponderant discipline (Cheng Y et al, 2017). Some studies investigated the characteristics and influencing factors of international cooperation (David H et al, 2018, Ling J, 2018). Some studies took a single country as research object and investigated its cooperation patterns and characteristics with different countries. Osorio N L et al (2017) investigated the patterns of research between Brazil and seven Latin American countries (Argentina, Chile, Colombia, Peru, Uruguay, and Venezuela); and between Brazil and countries of the major industrial democracies or G8 group: France, United States, United Kingdom, Russia, Germany, Japan, Italy and Canada. China has been actively engaged in international cooperation and has made some achievements. Research shows China was embedded in this top-layer of internationally co-authored publications (Leydesdorff L et al,2014).Also, Zhihui Z et al(2018) illustrates China's rising importance in scientific research and collaboration over the past 15 years by international co-authorship networks, China's co-authored and highly cited papers.

Scientometric scholars used to analyse international research cooperation by measuring papers and citation network. Because citation counting is more likely to generate errors and disputes than paper counting (Narin et al, 1999, Glanzel W, 2001), and the citation counting has the characteristics of retardance compared with paper counting, this paper adopts paper counting as the method to measure international cooperation(Braun T, 1989). Recent years, new approaches has been explored, for example Paul-Hus used 'Acknowledge' to discover the cooperation in scientific research (Paul-Hus, 2016). We suggest that authors' division of labor obtained from 'Author Contributions' can be used as a supplementary for scientometric research. Therefore, this research analyses the division of labor in high quality life science research of China from the perspectives of quantity and structure.

We parses the 'Author Contributions' text in life science publications in Nature Index 68, extracts the structure with author contribution semantics for the division of labor infromation, and carries on the quantitative statistics and structure analysis according to the nationality of each author. This research can interpret the division of labor in Chinese high quality life science research from the quantitative and structural, national and international perspectives.

Data and methods

Data

The data source is the publications from life science journals in Nature Index 68. The main reasons for choosing life science are as follows: First, life science field attachs more attention to the author contribution, thus the 'Author Contributions' is implemented earlier, and the writing format is relatively standard. Besides, life science play an important role in the total articles in Nature index 68. Therefore, we consider our data resource can reveal the division of labor in high-quality life science research.

Most journals in Nature Index 68 have a specific field classification. However, multidisciplinary journals (Nature, Science, Nature communications and PNAS) contain a lot of researches belonging to life science. Publications from them are re-classified according to their references: the field the most references belong to. So the data composes of two parts: publications from life science journals and that from multidisciplinary journals.

Publication's title, DOI and 'Author Contributions' are collected from the journals' website by crawler code (python). Other auxiliary information, including the author's name, author's country and literature type, is obtained from Web of Science. We firstly try to use DOI to combines these information. If a publications' DOI is missing or wrong, we use title as an alternative.

Literature type in this study is Article and Review. The data source and co-authorship papers between China and other countries are shown respectively in Figure 1 and Table 1.

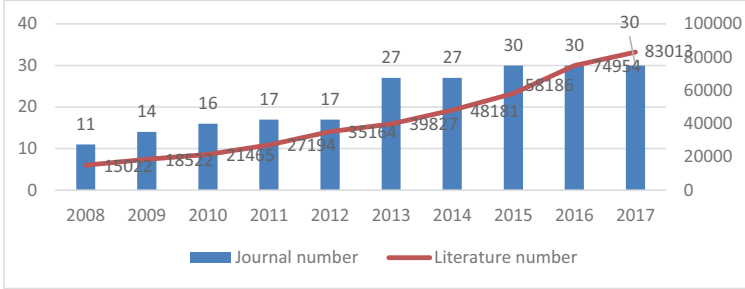


Figure 1.Total: 30 journals, 421528 publications

Table 1. Number of co-authorship papers between China and other countries in 2017

Country	*category*	*Co-authored papers number*	*Country*	*Category*	*Co-authored papers number*
USA	G7	881	Brazil	BRICS	7
Germany	G7	113	India	BRICS	13
UK	G7	121	Russia	BRICS	13
France	G7	56	South Africa	BRICS	5
Japan	G7	83			
Canada	G7	74			
Italy	G7	33			

When calculating the world share of national papers, full counts will cause the expansion of the total number of papers. On the contrary, fractional counts method is more accurate (Braunt, 1989). We consider every paper's credit is 1 and use fractional counting to count Chinses international cooperation status in this dataset. Each collaborating author of a paper equally shares one credit. We use the proportion of Chinese international cooperation credit by fractional counting as baseline, which investigate China's cooperation status from the macro level of paper quantity, in order to normalize the increase of international publication in China, as baseline for comparative analysis.

Methods

Categories of division of labor

Scholars take International Committee of Medical Journal Editors (ICMJE) standard to analyse the contributing factors of scientific research papers (Bates et al, 2004, Baerlocher et al, 2007). By means of questionnaire, interview and literature content analysis, Zhang divided the contributions of authors of journals in the field of life sciences into seven categories: Thought, Data, Experiment, Tool, Expression, Management and Other (Zhang, 2016). Our research is supposed to consider the significance of different authors' contributions not only at author level, but also at country level. In the division of labor analysis in country level, data acquisition and data analysis has different meanings, thus we divided data into Data Analysis and Sample or Data Collection. Besides, the other division of labor are not considered here due to the small number and messy content. Thus, the authors' division of labor is classified into seven categories: 1) Research Design, 2) Experiment Operation, 3) Data Analysis, 4) Sample or Data

Collection, 5) Tools and Technology, 6) Expression and 7) Management. The definition and representative rules of each categories are as follows.

Table 2. Seven categories of division of labor and definitions

Division of labor	*Definition*	*Representative Rules*
Research Design	Contributions to the thought, including the design of the structure of the paper or the experimental structure	'conceive-ideas'; 'design-research'
Experiment Operation	Contributions related to experiment implementation	'perform-experiments'; 'carry-purification'
Data Analysis	Analysis and interpretation of information, content and experimental results	'undertake-analysis'; 'analyse-data'
Sample or Data Collection	Acquisition and collection of data and samples	'collect-specimens'; 'obtain-data'
Tools and Technology	Refers to the equipment, reagents, models, storage environment and space provided to ensure the completion of the experiment	'provide-tools'; 'develop-techniques'
Expression	Contributions related to the expression of the research, including writing and revision	'draft-manuscript'; 'review-paper'
Management	A series of auxiliary activities in the research, including the arrangement of tasks between the personnel, communication in the experiment process, coordination and supervision process	'approve-manuscript'; 'supervise-experiment'

Data preprocessing method

Firstly, abbreviation of author name in 'Authors Contributions' text is converted into full name. Secondly, 'Authors Contributions' text is divided into sentences based on sentence separators (';', ',', etc.) and some fixed sentence structures as well as the POS information provided by Stanford POS Tagger. Then, according to different types of author contribution, the structure with author contribution semantics is extracted by syntactic parsing or direct matching. Finally, the extracted author contribution semantic structures are matched with the predefined rule libraries to obtain the corresponding category of division of labor.

(1) Establishment of rule libraries

Rule libraries are important tools storing the 'verb - object' structures or typical noun phrases representing division of labor. Newly collected 'verb - object' structures and noun phrases can be corresponded to pre-defined categories of division of labor accurately and quickly. The quality of rule libraries directly determines the effect of author contribution recognition.

Due to the two different kinds of contribution text format in the publications, we establish the rule libraries in the following ways. For natural language expression, 'verb - object' structure is extracted at first. Manually deleting meaningless structure, 'verb - object' structure frequency is calculated. If the frequency beyond 5, the 'verb - object' structure is a rule with author contribution semantics. As a result, the rule library contains a total of 4063 rules derived from 'verb - object' structure, covering 92.9% of the author contribution semantic structures.

Similarly for direct corresponding, 'Author Contributions' sentences are divided according to the author contribution semantics. Manually deleting meaningless structure, the frequency of typical noun phrase is calculated. If the frequency beyond 5, the typical noun phrase is a rule with author contribution semantics. As a result, the rule library contains a total of 1364 rules derived from typical noun phrase, covering 55.5% of the author contribution semantic structures.

(2) Extraction of author contribution semantic structure

For the contribution in the form of natural language, the semantic structure extraction mainly depends on the POS information of Stanford POS Tagger and the dependency grammar

information of Stanford Dependency Parser. In this research, complex sentence is divided into simple sentences to increase recognition efficiency. For the directly corresponding contribution, such as a piece of contribution in PLOS Biology in 2017: *Annick Sawala: Conceptualization, Formal analysis, study, Methodology, the Visualization, Writing - the original*, it is recognized according to the predefined rule library of typical noun structure.

Data analysis method

(1) Calculation of division of labor

Author contributions are divided into seven categories: 1) Research Design, 2) Experiment Operation, 3) Data Analysis, 4) Sample or Data Collection, 5) Tools and Technology, 6) Expression and 7) Management. If one category of division of labor exists in one article (C_i, $i \in \{1,2,3,4,5,6,7\}$), record as 1, otherwise 0.

Count the number of authors who undertake one category of division of labor as n, then the relative value of one author's this category of division of labor (P_i) is C_i/n .

(2) Quantity analysis

In order to normalize the influence of the national scientific research growth trend and obtain the temporal sequence proportion of the seven kinds of division in China in the world, we divide the value of the seven categories of division of labor in China in 2008-2017 by the value in the world every year. The proportion of the world and the international ranks aim to study the change of China in the world in the past decade on seven categories of division of labor.

(3) Structure analysis

In order to analyse the changes of contribution structure of Chinese authors in the past decade, two time windows (2008 and 2017) are selected to explore the changes in the proportion of the seven categories of division of labor in the total contributions of Chinese authors in one year.

Comparison of China with G7 and BRICS countries (excluding China) in terms of quantity and structure is conducted. Firstly, the distribution of seven categories of division of labor in 2017 was selected to compare the similarity between China and these countries. We regard each country as a seven-dimensional vector, and each dimension is the proportion of one category of division of labor. The cosine function is applied as the similarity calculation method.

To further explore the co-authorship model between China and different countries, co-authored papers of China with G7 and BRICS countries (excluding China) were selected respectively to analyse the co-authorship ratio between two countries in the seven categories of division of labor.

Result

Quantity analysis

(1) Proportion of Chinese seven kinds of divisions from 2008 to 2017

Figure 3 shows the time series analysis of Chinese proportion of seven categories of division of labor in the world in recent ten years. In general all categories is on the rise, from about 2% to about 6%, indicating that Chinese position in each division of labor in high-quality life science research is on the rise. Contributions of Research Design, Experiment Operation, Data Analysis, Expression and Tools and Technology are on a steady upward trend, while the contributions of Management and Sample or Data collection are relatively fluctuating due to the small data, possibly.

On the whole, the variation trend of seven categories of division of labor was basically consistent with baseline. The division of labor among Chinese authors are relatively balanced. These seven categories of division of labor could roughly divided into three classifications. The first is that the proportion is always higher than baseline since 2008, indicating that China has always been in the leadership in these divisions of labor, including Experiment Operation and

Management. While the second are those whose proportions are below baseline in 2008, but gradually higher than baseline as the year goes by, such as Research Design and Data Analysis. Among them, Data Analysis, although lower than baseline in 2008, is already higher than baseline in the following year. However, Research Design, until year 2017, exceeds baseline for the first time. It is inferred that as year goes by, an increasing number of Chinese researchers are inclined to undertake these two kinds of division of labor. The third which contains Sample or Data collection, Expression and Tools and Technology, have been in a state of continuous fluctuation corresponding with baseline during 2008 to 2017.

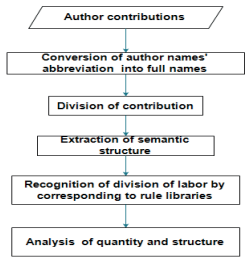


Figure 2. Overall technical flow chart

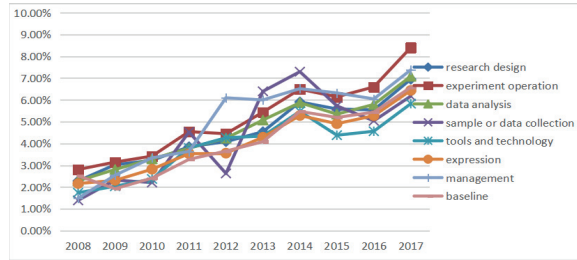


Figure 3. Proportion of Chinese seven categories of division of labor from 2008 to 2017

(2) Comparison of Chinese ranks of seven divisions of labor in 2008 and 2017

Table 1 shows that in 2008, except for Management, each division of labor and baseline of China are ranked the 8th among G7+BRICS countries. Besides, the 7 countries ahead of China were exactly the seven countries of G7. While in 2017, Chinese Experiment Operation was ranked the 3rd, only behind USA and Germany, which shows that in the field of life science, China has been playing an increasingly important position in Experiment Operation. Considering the characteristics of life science, where experiment plays an important role, not only in the total quantity of publications, but also in the relative important division of labor, Chinese scientists have been undertaking an increasingly important role.

Structure analysis

(1) Comparison of Chinses seven categories of division of labor in 2008 and 2017

Figure 4 shows that among the seven categories of division of labor, Management has increased greatly, from 10.49% in 2008 to 15.29% in 2017. In addition, Data Analysis has improved to a certain extent, Tools and Technology has remained basically stable, while Research Design, Experiment Operation, Data Analysis and Expression have decreased.

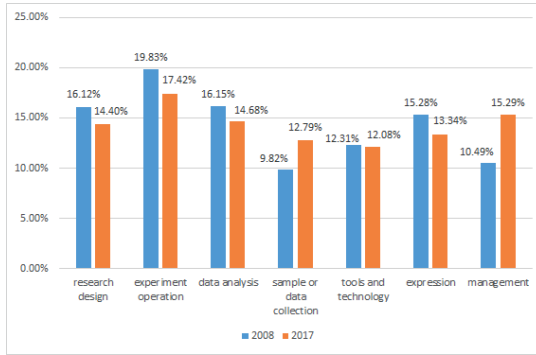


Figure 4. Seven categories of division of labor between 2008 and 2017 in Chin

Table 3. The proportion and ranks between China and 11 countries in 2008

Country	*Percent of contributions*								*Rank*							
	Research Design	*Experiment Operation*	*Data Analysis*	*Sample or Data Collection*	*Tools and Technology*	*Expression*	*Manage ment*	*Baseline*	*Research Design*	*Experiment Operation*	*Data Analysis*	*Sample or Data Collection*	*Tools and Technology*	*Expression*	*Manage ment*	*Baseline*
Peoples R China	2.28%	2.80%	2.29%	1.39%	1.74%	2.16%	1.48%	2.51%	8	8	8	8	8	8	7	8
USA	41.52%	39.62%	39.34%	27.48%	39.15%	42.76%	37.19%	37.80%	1	1	1	1	1	1	1	1
Germany	7.25%	7.92%	7.85%	6.67%	8.10%	7.42%	8.71%	7.59%	3	2	3	4	2	3	3	3
UK	8.08%	7.04%	8.95%	7.45%	7.86%	8.90%	10.65%	10.43%	2	3	2	2	3	2	2	2
France	4.23%	4.79%	4.73%	3.41%	6.42%	4.07%	4.40%	4.83%	4	4	4	6	4	4	6	5
Japan	4.18%	4.14%	4.19%	7.29%	6.15%	3.26%	4.73%	6.03%	5	5	6	3	5	6	5	4
Canada	3.69%	3.58%	4.25%	4.75%	5.14%	3.71%	6.44%	4.77%	6	6	5	5	6	5	4	6
Italy	3.31%	3.35%	3.04%	3.33%	2.84%	2.99%	1.48%	2.53%	7	7	7	7	7	7	8	7
Brazil	0.33%	0.38%	0.40%	0.00%	0.05%	0.28%	0.00%	0.15%	9	9	9	9	12	9	10	9
India	0.09%	0.12%	0.12%	0.00%	0.11%	0.09%	0.00%	0.05%	12	12	12	9	11	12	10	12
Russia	0.24%	0.30%	0.34%	0.00%	0.23%	0.23%	0.00%	0.06%	11	10	10	9	10	11	10	11
South Africa	0.28%	0.23%	0.32%	0.00%	0.26%	0.24%	0.29%	0.01%	10	11	11	9	9	10	9	10

Table 4. The proportion and ranks between China and 11 countries in 2017

Country	*Percent of contributions*								*Rank*							
	Research Design	*Experiment Operation*	*Data Analysis*	*Sample or Data Collection*	*Tools and Technology*	*Expression*	*Manage ment*	*Baseline*	*Research Design*	*Experiment Operation*	*Data Analysis*	*Sample or Data Collection*	*Tools and Technology*	*Expression*	*Manage ment*	*Baseline*
Peoples R China	6.95%	8.40%	7.08%	6.17%	5.83%	6.44%	7.38%	6.58%	4	3	4	5	4	4	4	4
USA	36.97%	34.99%	35.59%	28.47%	35.74%	37.52%	34.97%	35.76%	1	1	1	1	1	1	1	1
Germany	8.60%	9.17%	9.25%	8.08%	10.24%	8.85%	8.88%	9.75%	2	2	2	3	2	2	2	2
UK	8.30%	7.55%	8.35%	11.25%	7.29%	8.62%	8.73%	7.74%	3	4	3	2	3	3	3	3
France	5.12%	5.43%	5.13%	6.81%	5.41%	4.94%	5.72%	5.55%	5	5	5	4	5	5	5	5
Japan	2.82%	3.07%	2.71%	2.00%	4.55%	2.75%	3.10%	3.17%	7	7	7	8	6	7	6	7
Canada	3.63%	3.33%	3.53%	3.36%	3.41%	3.50%	2.91%	3.76%	6	6	6	6	7	6	7	6
Italy	1.79%	2.01%	1.90%	2.52%	2.17%	1.89%	2.16%	2.05%	8	8	8	7	8	8	8	8
Brazil	0.34%	0.33%	0.39%	0.33%	0.29%	0.27%	0.33%	0.48%	10	10	10	12	10	10	9	10
India	0.40%	0.50%	0.55%	0.81%	0.40%	0.44%	0.30%	0.55%	9	9	9	9	9	9	10	9
Russia	0.25%	0.24%	0.27%	0.39%	0.27%	0.25%	0.18%	0.32%	11	11	11	11	11	11	11	11
South Africa	0.18%	0.15%	0.18%	0.78%	0.25%	0.21%	0.15%	0.19%	12	12	12	10	12	12	12	12

(2) Structure comparison of division of labor between China and 11 countries in 2017

Calculating the cosine similarity, the country most similar to China is Brazil, one of the BRICS. Except the proportion of Chinese Experiment Operation is slightly higher than Brazil while the proportions of Brazil's Data Analysis and Research Design are slightly higher than China, two countries are almost identical. The following similar countries are USA and Germany, whose cosine similarity with China are more than 99%. From the radar plots, except Tools and Technology in USA and Germany have higher proportions, these four countries have relatively balanced distributions of each division of labor on the whole.

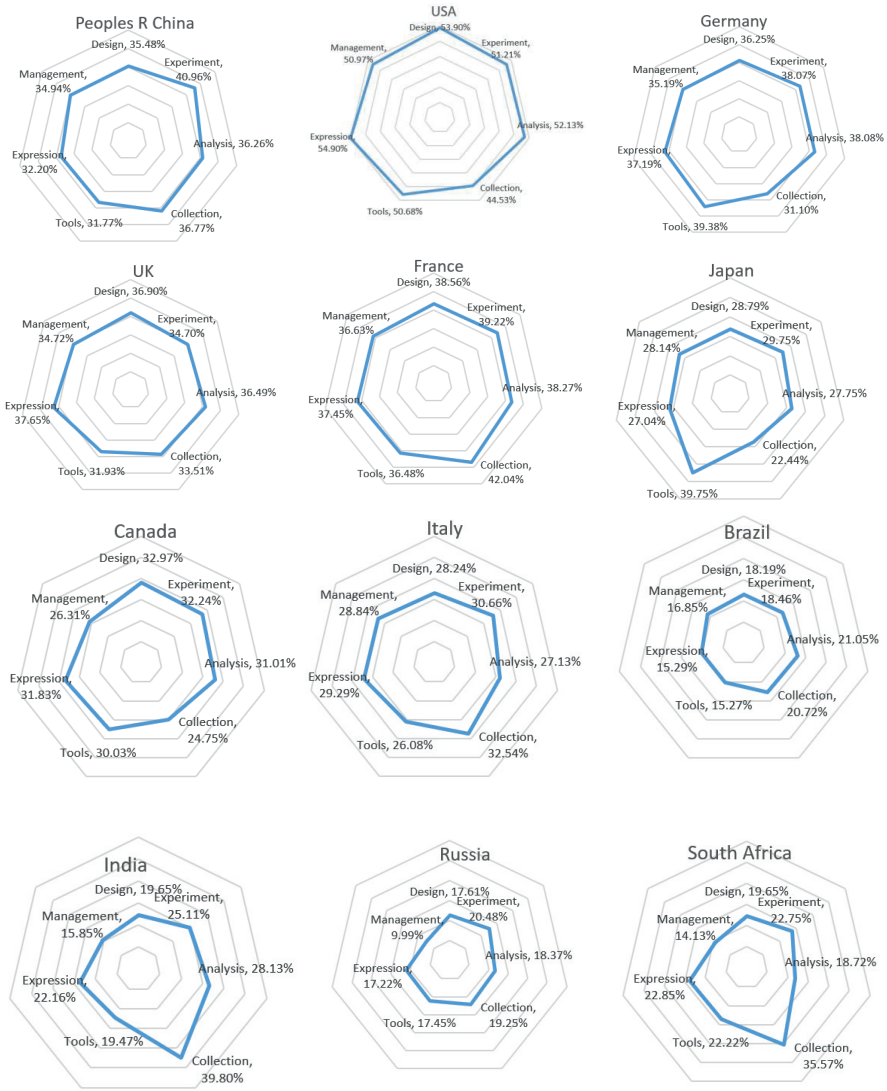


Figure 5. Structure of division of labor of 12 countries

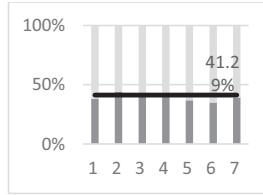
Table 5. The similarity between China and 11 countries

Country		*Similarity*	*Rank*
USA	G7	99.36%	5
Germany	G7	99.34%	6
UK	G7	99.54%	4
France	G7	99.78%	1
Japan	G7	97.79%	9
Canada	G7	99.24%	7
Italy	G7	99.75%	2
Brazil	BRICS	99.58%	3
India	BRICS	96.33%	11
Russia	BRICS	98.63%	8
South Africa	BRICS	96.44%	10

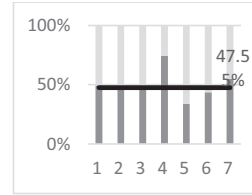
Except for South Africa, the rest countries have reached more than 90% cosine similarity with China, which indicates that most countries' contribution distributions are relatively balanced. South Africa's division of labor are extremely unbalanced, Sample or Data Collection is South Africa's main author contribution, considering the amount of publications, the country's scientific research in life science field is still in its infancy.

(3) Analysis of co-authorship models between China and other countries

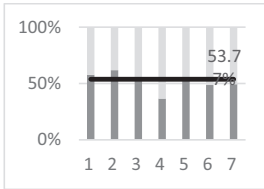
Number	***Division of labor***
1	research design
2	experiment operation
3	data analysis
4	sample or data collection
5	tools and technology
6	expression
7	management



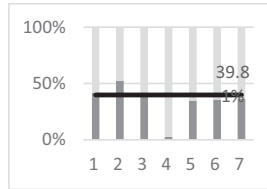
USA



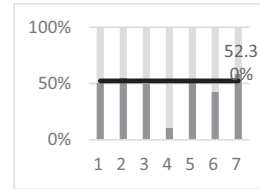
Germany



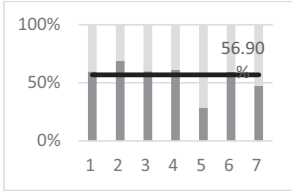
UK



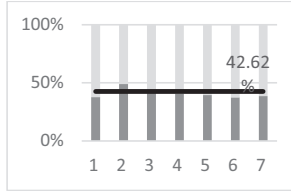
France



Italy



Japan



Canada

Figure 6. Co-authorship model between China and 7 countries

In Chinese scientific cooperation with the United States, whether from seven categories of division of labor or baseline, the United States is always in the relative leading position. Specific analysis of each division of labor shows that the two countries have a balanced pattern of scientific cooperation (basically consistent with baseline). In Experiment Operation and Sample or Data Collection, China slightly exceeds baseline, while in other division of labor, the United States slightly exceeds China. It can be concluded that Chinese scientists are mainly engaged in data work among the two countries' cooperation. The co-authorship model of Canada and China is similar to that of United States and China.

In the cooperation between China and other countries, compared with baseline, Sample or Data Collection of China is significantly higher than Germany, which is contrary to China with UK, France and Italy. In the co-authored papers with these countries, China occupies a significantly lower proportion in Sample or Data Collection. For Tools and Technology, the proportion of Germany and Japan are significantly higher than China, which indicates that in the cooperation between China and these two countries, these two countries tend to provide more tools and technology support.

Since the numbers of co-authored papers between BRICS and China are very small, the conclusion is greatly affected by small sample and are not statistically significant, so the four countries are not analysed here.

Discussion and conclusions

In the past decade, for high-quality life science research, contributions of all Chinese authors have been improved from both the numerical dimension and the international rank dimension, of which Experiment Operation plays an increasingly important role in the division of labor. From the change of domestic division of labor structure, China shifted the focus of division of labor to Management and Data Analysis in the past decade. From the comparative analysis of the structure of international division of labor, China and all countries except South Africa have similar roles in division of labor, among which Brazil, USA and Germany are the most similar. From the perspective of co-authored papers in China and other countries, co-authored models in China, USA and Canada are relatively balanced. And the differences between co-authorship models of China and other countries are mainly reflected on the contributions of Sample or Data Collection and Tools and Technology.

Basu et al. analysed China's scientific and technological status from qualitative, quantitative and supplementary perspectives. He believed that China was gradually restoring its leading position in science and technology in recent years (Basu et al, 2018). This study can reach the similarly conclusions according to the analysis of division of labor. This study suggests that in the field of life science, China's scientific research has developed rapidly and its role in experimental operation has been continuously improved, but China still relies on Germany and Japan for technology and tool support.

This study has the following limitations: First, the recognition effect is greatly affected by the rule libraries, and the coverage of rule libraries needs to be further improved. Second, the data source of this research is articles and reviews in life science field of Nature index 68. Compared with other scientometric paper, the amount of data of this research is relatively small, so the conclusion may have some deviation. For example, in China and other countries' co-authored mode analysis, we can not draw a reliable conclusion for China and other BRICS countries due to the small quantity of co-authorship papers.

In the future study, we will further expand the rule libraries and select a larger range of data source. Furthermore, we will try to integrate existing indicators and build our own model in order to conduct deeper analysis of the co-authorship model among different countries.

Acknowledgments

Thanks Liying Yang, professor of National Science Library, Chinese Academy of Sciences, for advice and suggestion.

References

Baerlocher, M. O., Newton, M., Gautam, T., et al. (2007). The meaning of author order in medical research. *Journal of Investigative Medicine*, 55 (4), 174-180.

Basu, A., Foland, P., Holdridge, G., et al. (2018). China's rising leadership in science and technology: quantitative and qualitative indicators. *Scientometrics*, 117, 249-269.

Bates, T., Ani, A., Marusi, M., et al. (2004). Authorship criteria and disclosure of contributions: comparison of 3 general medical journals with different author contribution forms. *JAMA*, 292(1), 86-88.

Cheng, Y., Liu, Y., Wang, W. (2015). Empirical research on international S&T cooperation promoting the Annual Conference of China Soft Science. Beijing

Hsiehchen, D., Espinoza M., Hsieh, A. (2018). Evolution of collaboration and optimization of impact: self-organization in multinational research. *Scientometrics*, 117(1), 391-407.

Feeser, V. R., Simon, J. R. (2014). The ethical assignment of authorship in scientific publications: Issues and guidelines. *Academic Emergency Medicine*, 15(10), 963-969.

Larivière, V., Desrochers, N., Macaluso, B., et al. (2016). Contributorship and division of labor in knowledge production. *Social Studies of Science*, 46(3), 417-435.

Leydesdorff, L., Wagner, C. S., Bornmann, L. (2014). The European Union, China, and the United States in the top-1% and top-10% layers of most-frequently cited publications: Competition and collaborations. *Journal of Informetrics*, 8(3), 606-617.

Jiang, L., Zhu, N., Yang, Z., et al. (2018). The relationships between distance factors and international collaborative research outcomes: A bibliometric examination. *Journal of Informetrics*, 12(3), 618-630.

Osorio, N. L. & Otieno, A.W. (2007). International cooperation of Brazilian research engineers: Patterns of collaboration based on a survey of the literature. *2007 37th Annual Frontiers In Education Conference - Global Engineering: Knowledge Without Borders, Opportunities Without Passports*. EUA: Milwaukee, WI.

Paul-Hus, A., Mongeon, P., Sainte-Marie, M., et al. (2016). The sum of it all: Revealing collaboration patterns by combining authorship and acknowledgements. *Journal of Informetrics*, 11(1), 80-87.

Price, D. D. (1981). Multiple authorship. *Science*, 212(4498), 986.

Ribeiro, L. C., Rapini, M. S., Silva, L. A., et al. (2017). Growth patterns of the network of international collaboration in science. *Scientometrics*, 114(1), 159-179.

Yank, V. (1999). Disclosure of researcher contributions: a study of original research articles in the Lancet. *Annals of internal medicine*, 130(8), 661-670.

Zhang S. (2016). Research on Optimization of Author Contribution Weight Algorithm, Ph.D. Dissertation. Beijing: University of Chinese Academy of Sciences.

Zhang, Z., Rollins, J. E., Lipitakis, E. (2018). China's emerging centrality in the contemporary international scientific collaboration network. *Scientometrics,* 116(2), 1075-1091.

A Multi-Dimensional Observation Framework of Retracted Publications

Junpeng Yuan[1], Lingzi Feng[2], and Liying Yang[3]

[1] *yuanjp@mail.las.ac.cn*

[2] *fenglingzi@mail.las.ac.cn*

[3] *yangly@mail.las.ac.cn*
National Science Library, Chinese Academy of Sciences, Beijing (China)
Department of Library, Information and Archives Management, School of Economics and Management, University of Chinese Academy of Sciences, Beijing (China)

Abstract

In order to describe the impact of the retracted publications on the scientific community more comprehensively, this paper introduces the impact of retracted publications on technology, establishes the Multi-Dimensional Observation Framework of Retracted Publications from the five aspects, including Scientific Impact, Scientists Impact, Technology Impact, Funding Impact and Altmetric Impact, and propose the use of multi-dimension bubble diagram and sanky diagram to show the multi-dimensional data. This paper takes all the retracted publications from the journal *Cell* as an example, using the observation framework to analyze them, and describes the impact of these papers from multiple dimensions. The results show that compared with the single dimension, the observation framework reveals the impact characteristics of retracted publications abundantly to some extent, and also expands the dimension of retracted publications' impact.

Introduction

Retracted publications, i.e. papers that have been reported by members of the Scientific Community or applied by authors, and investigated then found problems such as scientific misconduct or scientific errors, thus retracted by publishing journals through formal retraction notice because of the negative impact it may cause on the development of science (Fan & Zhang, 2014). There have been abundant researches on the phenomenon of retractions. Some scholars (Zhang, Ding, & Liu, 2018) classify the existing research into three categories: feature, impact and reason. Researches on feature are mainly about time (publication time or retraction time), authors' geographical distribution (country or region), journal, involved research entity (author or institute) and retraction time lag. Researches on impact are mainly about description of cited status(number of citations, its changes and attitude of citers etc.), citing reasons and influence factors, additional impact and countermeasures. Researches on reason are mainly about classification and statistics of retraction reasons, feature differences under different retraction reasons and specific retraction reason analysis. Our research could be classified as the category impact. And the meaning of the word 'impact' in this article is more of characteristics of retracted publications, which is associated with data like citation, download etc..

In researches that studies the impact of retracted publications, scholars analysed citation numbers, citation numbers before and after retraction and sentiment of citation sentences, at present, it is believed that the citation rate of retracted publications generally decreases after retraction, but there are still a large number of citations, and most of them are still non-negative citations (Wright, 1991; Shuai et al., 2017; Bar-Ilan & Halevi, 2017; Luwel et al., 2018;). Retracted publications also involves the impact on funding, such as directly affecting the funded grant (Stern et al., 2014), and affecting the funding application of involved discipline (Azoulay et al., 2012). Retracted publications also influence citations of authors and their collaborators

(Mongeon & Larivière, 2016), such as influence scholars at difference stages of career differently (Jin et al., 2013; Azoulay et al., 2017), or influence discipline differently (Shuai et al., 2017). Scholars have also explored the performance of retracted publications on Altmetric indicators (Bar-Ilan & Halevi, 2018). However, researches that study the impact of retraction on technology and various aspects comprehensively are little known.

This study proposes an observation framework for retracted publications, it considers the citations received from patents, which has drawn little attention at present, whether the original conclusions influence the development of scientific development, and combines data from different sources, such as Web of Science citation, F1000 Prime (peer review recommendation platform) recommendations, number of funding involved, and scores of comprehensive indicators of Altmetric. We are trying to describe the impact that has been produced by retracted publications comprehensively from five dimensions, namely Scientific Impact, Scientists Impact, Technology Impact, Funding Impact, and Altmetric Impact. In order to display and analyze the result in a significant way, we choose not to integrate the result of five dimensions in one number but propose the use of multi-dimensional bubble diagram and sanky diagram to show the data in multiple dimensions.

Methods and Data

This study proposes an observation framework for describing the impact of the retracted publication produced from multiple dimensions. This framework (Fig.1) includes five dimensions mentioned above. The contents and data of the framework are as follows:

Scientific Impact, i.e. the impact of retracted publications on scientific development. If the conclusions of the retracted publications are wrong and unreliable, the dissemination of the papers with wrong conclusions will inevitably have a negative impact on those researches relying on erroneous conclusions and even on the development of science and technology as a whole. Bar-Ilan & Halevi (2018) proposed a classification of retraction reasons based on whether there is scientific distortion, we decide to extract the fact whether there is the wrong conclusion as a dimension in our framework and retain other factors from retraction notice like academic misconduct in the data field of retraction reasons. In terms of indicators, this dimension is composed of the index judging whether the conclusion of this retracted publication affects the development of the scientific community. In terms of data, this dimension only has two value 1 or 0, the value is given by the judgment of retraction reason extracted by Retraction Watch Database (http://retractiondatabase.org) and the interpretation of specific retraction notice content. The rules are as follows: if the retraction notice states that the original conclusion is affected, the result of this research cannot be reproduced, the result is not consistent with the original conclusion, or the retraction notice does not expressly indicate whether the conclusion is correct or wrong, or there is a dispute between authors about original conclusion, the value is 1. If the retraction notice expressly states that the conclusion is unaffected, then the value is 0.

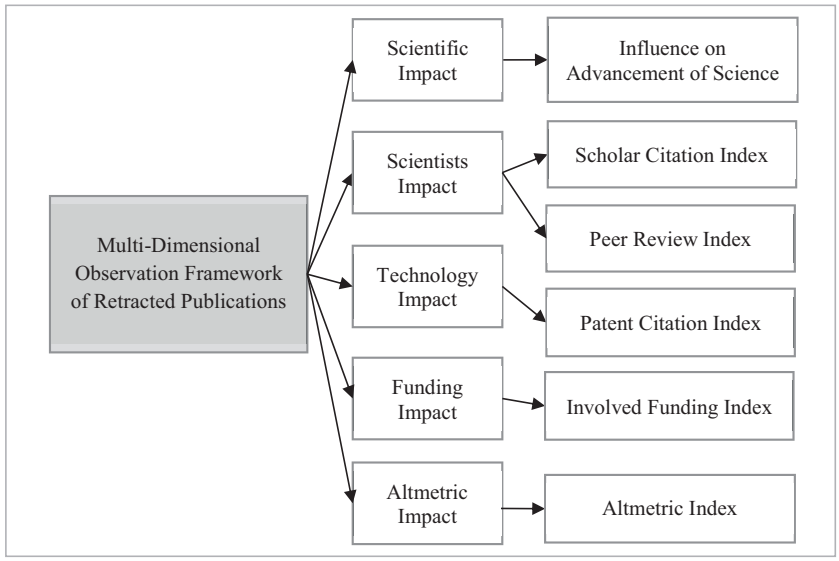


Figure 1. A Multi-Dimensional Observation Framework of Retracted Publications

Scientists Impact, i.e. the impact of retracted publications on scientists, which could be revealed by citing and recommending behaviors. Because the citing behavior is comprised of the reading, comprehension, and attitude of scientists who list a publication as reference or recommend it, theoretically, the more scientists cite and recommend a retracted publication, the greater the impact on scientists. In terms of indicators, this dimension is composed of scholar citations and the number of stars recommended by peer review platforms. In terms of data, scholar citation comes from Clarivate based platform Web of Science, the value is the number of citations, and recommendation data for peer review are from F1000 Prime[1], the value is the sum of recommendation stars. The final dimension value SI_i is determined by formula (1), in which C_i represents the citation number of paper i, and R_i is the total number of recommendation stars that paper i received.

$$SI_i = C_i \times \left(\frac{1}{2} + \frac{1}{1+e^{-R_i}}\right) \quad (1)$$

Based on citation, the formula enlarges the citation value to a certain extent by taking the recommended star value as the expansion coefficient. The coefficient is calculated using the Sigmoid function (i.e. growth curve), which maps the values of infinite intervals to finite intervals. We take the sum of stars R_i as the input of the sigmoid function. Since the range of R_i is $[0,+\infty)$, the range of Sigmoid function is [0.5, 1). In order to get the effective interval, the expansion coefficient is shifted to [1,1.5) by translating the whole function upward by 0.5. After testing, this fusion method can not only ensure that the recommended star number plays an important role, but also ensure the rationality of the effect. Because the range of the expansion coefficient is limited, its influence on the expansion of citation is also within a reasonable range, which guarantees the priority of citation to a certain extent. For example, when the citations of two papers vary greatly in different scales (the difference is more than 1.5 times), even if the articles with less citations have more stars, the other one will not exceed the scores of the articles with higher citations after expansion. When the citations of the two articles are similar and it is not easy to distinguish the impact, the expansion coefficient determined by the recommendation

[1] This platform recommends best scientific articles from biological and medical fields by global leading experts, they rate articles as 'Good', 'Very Good' or 'Exceptional', which equivalent to scores of 1, 2 or 3 stars respectively.

star number can play a good role in adjusting. Generally speaking, the formula can effectively and reasonably integrate the results of two incomparable values of citation and recommendation star number.

Technology Impact, i.e. the possible impact of retracted publications on technology, because there is a certain possibility that if a patent cites a retracted publication, the patent itself would be affected. For example, affected by wrong conclusions if the patent is based on it, or directly or indirectly influenced by academic misconduct. The more times the retracted publication is cited by patents, the greater the possibility and scope of the impact would be. In terms of indicator, this dimension is composed of the number of times that retracted publications are cited by patents (considering that patents that have larger patent family is more important, we decided to count patents as how many times it was cited by all patents separately, instead of count the number of patents in one patent family as 1 citation). In terms of data, patent citation data are from lens.org[2], the value is the number of patents that cite a certain retracted publication.

Funding Impact, i.e. the possible impact of retracted publications on funding. The allocated grant money may be directly or indirectly wasted by academic misconduct, wrong experiment operation and so on, we wonder to what extent this kind of impact is, and assume that the more the projects supported and the larger the amount involved, also the greater the theoretical impact. In terms of indicator and data, this dimension is the number of projects or grants involved in the Funding field of a paper in the Web of Science database. Stern et al. (2014) used NIH funding data to estimate the approximate amount of funding for each retracted publication, but because a high proportion of the paper funding information in this case is incomplete and the amount cannot be collected, the number of projects is used instead.

Altmetric Impact is a reflection of the impact of retracted publications being read, used and commented on the Internet, such as the number of social media mentions, academic sharing counts and so on. The attention and discussion over scholar works may be a supplement to the traditional citation. Sometimes the Altmetric index like usage and comment may reveal the impact of retraction more directly than citation, because people may tend to discuss and download other than cite a retracted publication. In terms of indicator and data, Altmetric.com proposes an Altmetric Score that combines multiple Altmetric indicators. This index gives a comprehensive score based on the data of each paper mentioned on Twitter and Facebook, and the number of times read and shared on Mendeley, etc. At present, this indicator has been recognized and promoted to a certain extent, so this study uses the API of Altmetric.com to obtain the Altmetric Score of each retracted publication according to DOI or PMID of each paper, and chooses this score as the impact score of Altmetric Impact dimension.

This study takes all the retracted papers published by *Cell* as an example. The bibliography data and reasons for retraction are obtained from Retraction Watch Database. In terms of retrieval scheme, 'Cell' is used as keyword in Journal field. 'Research Articles' or 'Review Articles' are selected as the literature type. Nature of Notice we chose 'Retraction', those retrieved papers were formally issued a retraction notification. A total of 41 records were found. After verification on *Cell*'s official website (https://www.cell.com), ScienceDirect and PubMed, we excluded two records that were actually Correction and mislabelled Retraction, and finally obtained 39 records. This sample scale may be a little bit small, but we want to test our observation framework preliminarily on a complete and relatively famous journal with collectable data. In this case, data were acquired in December 2018, and all the data in our dataset were finally updated on January 1, 2019.

[2] lens.org is a data platform based on global patent citation scientific papers jointly developed by Cambia and Queensland University of Technology in Australia. In collaboration with Crossref and the National Medical Library of the United States, lens.org maps the PMID number and DOI number of patents of the United States Patent and Trademark Office, the World Intellectual Property Organization, the European Patent Office and the Chinese National Intellectual Property Administration and their cited scientific papers. The data platform has been used by Nature Index to evaluate the impact of basic research on technological innovation

Results and Discussion

Overview of Retracted Publications of Cell

In terms of retraction time, the publication and retraction time of the 39 papers retracted by *Cell* are shown in Fig. 2. The horizontal axis of the graph is the year and the vertical axis is the number of papers. It can be seen that there are relative time lags between publication and retraction of the papers. The retraction time lag distribution of *Cell* is shown in Fig. 3. It can be calculated from the data that the average retraction time lag of this journal is 3 years. As can be seen from the graph, most of the papers retracted in this journal were within 1-2 years, and only one paper was retracted more than 10 years after publication.

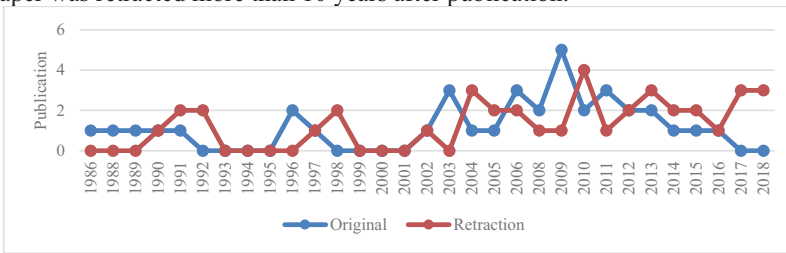


Figure 2. Publication and Retraction Year of *Cell* Retracted Publications

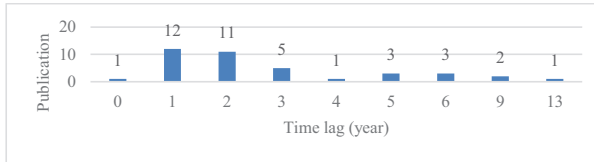


Figure 3. Time lag(year) of *Cell* Retracted Publications

In terms of retraction reasons, we have investigated and sorted out the retraction reason in Retraction Watch Database. According to the retraction guidelines published by COPE(Committee on Publishing Ethics)(Wager et al., 2009) and the retraction reason classification by Bar-Ilan & Halevi (2018), we have classified retraction reason occurred in Retraction Watch into 7 categories: scientific errors, academic misconduct, copyright issues, management mistakes, ethical violation, authorship issues and others. The result is shown in Table 1, 51.3% (20) papers were retracted due to scientific errors, 46.2% (18) papers were related to academic misconduct, while one paper is retracted because the content of the retraction notice was unclear, so we classified them as category Others.

Table 1. Retraction Reasons for *Cell* Retracted Publications

Retraction Reason	*Count*
Scientific Errors	20
Academic Misconduct	18
Others	1

In terms of whether there is wrong conclusion that affects the development of scientific community, as shown in Fig. 4, according to Retraction Watch Database's annotation of the retraction reasons for the retracted publication and our analysis of the retraction notice for each retracted paper, we found that 84.6% (33) of the retracted publications have the conclusions that may affect the development of scientific community, and only 15.4% (6) of the papers indicated that the conclusions would not be affected even if they are retracted.

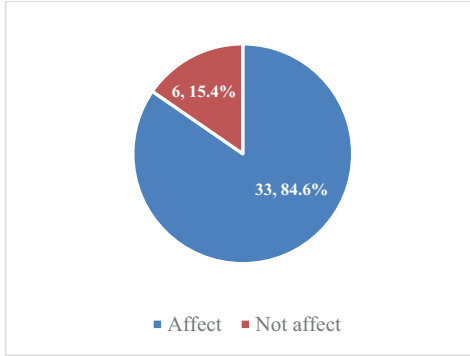


Figure 4. Proportion of the wrong conclusions that affect the development of scientific community of *Cell* Retracted Publications

Impact of Cell Retracted Publications Applying Multi-dimensional Observation Framework of Retracted Publications

Then we will apply the proposed observation framework of retracted publication to analyze the case. In order to describe and demonstrate the impact of the retracted publications from multiple dimensions, we believe that two visual observation maps can be used. One is the Sanky diagram shown in Fig.5, which divides the data results of the five dimensions into five columns and arranges dimension value in descending order. Each paper is connected by a line among dimensions. From such graph one can clearly see the values of each dimension, and intuitively observe the consistency and change of ranking of every dimension. The other one is a two-dimensional bubble diagram showing multi-dimensional data as Fig.6, which presents the results in two-dimensional space, with Scientists Impact on the horizontal axis and Technology Impact on the vertical axis. The bubble size is determined by Altmetric Impact, and the higher the value, the larger the bubble; the bubble color is determined by Funding Impact, the darker the bubble color, the more projects involved. As for Scientific Impact dimension, if the original conclusion of the retracted publication is wrong, it may have a negative impact on the scientific development process of the scientific community, then there is a black circle outside the bubble. From the two-dimensional bubble diagram one can clearly see the distribution and characteristics of the paper. In Fig.6, the order of papers numbered 1 to 39 is in descending order of citation. As can be seen from the graph, the results obtained from the observation framework are different from those only obtained from citations. As the dimension of the observation is expanded, the influence of the retracted publication can be shown more thoroughly and accurately.

In addition, Pearson correlation coefficients between the dimensions in this framework are calculated to analyze the correlation among the dimensions to help understand the results. The correlation coefficients are shown in Table 2. Since the Scientific Impact dimension has only two values 1 and 0, the values do not change much, so we did not calculate the correlation between this dimension and other dimensions.

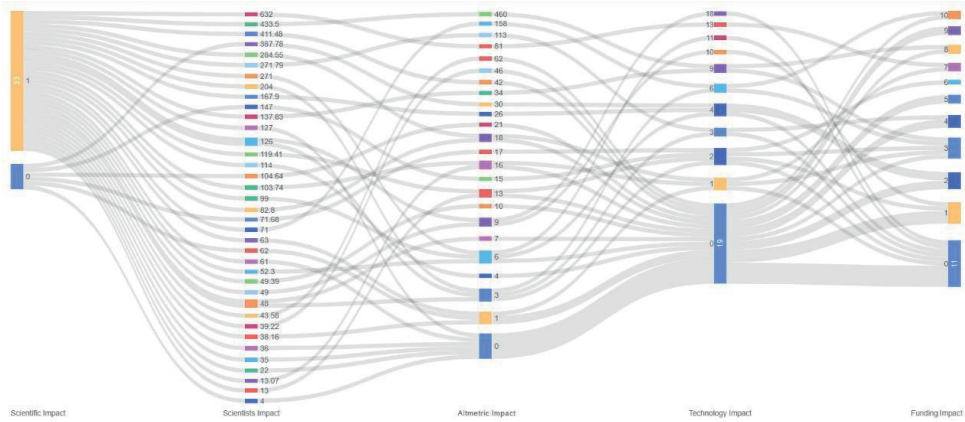


Figure 5. Impact of *Cell* Retracted Publications Applying Multi-Dimensional Observation Framework of Retracted Publications, Sanky Diagram

Each column of sanky diagram represents a dimension, each complete rectangle and each color in the column represents a certain numeric value, with the number of papers inside the rectangle (height of rectangle also represents the number of paper) and numeric **values** outside the rectangle. Each paper is connected by a line among 5 dimensions. In this figure, each column from left to right is Scientific Impact, Scientists Impact, Altmetric Impact, Technology Impact, and Funding Impact.

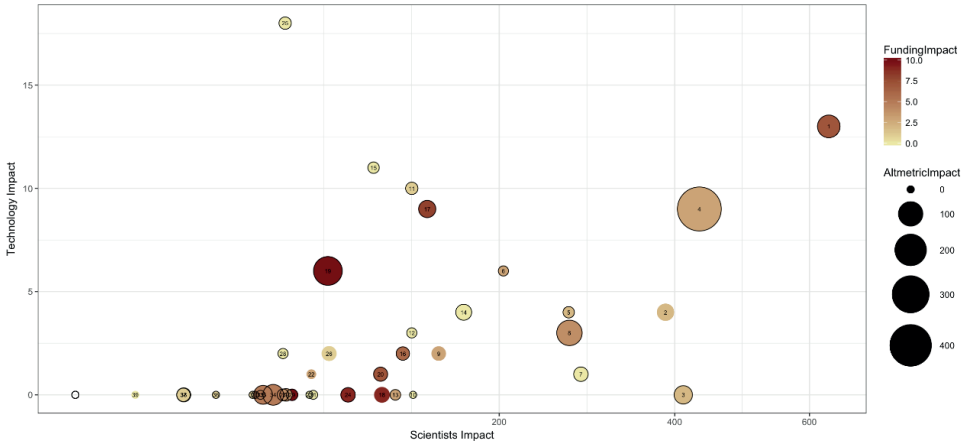


Figure 6. Impact of *Cell* Retracted Publications Applying Multiple Dimension Observation Framework of Retracted Publications, Multi-Dimension Bubble Diagram

The order of papers numbered 1 to 39 is in descending order of citation numbers

According to Fig. 5 and Fig. 6, and considering five dimensions, among the 39 retracted publications from *Cell*, papers 1 and 4 have relatively the highest impact. At the same time, both papers have relatively the highest Scientists Impact, indicating that the papers with the greatest impact among scientists have a high impact in all dimensions.

Looking at the five dimensions separately, Paper 1 was ranked top in Scientists Impact dimension, and the dimension value is 632. The result in Scientific Impact dimension is 1, it represents that there are erroneous conclusions that may affect the process of scientific development. At the same time, the paper has relatively high values in other dimensions,

Technology Impact dimension value is 13, ranked 2nd, Altmetric Impact dimension value is 81, ranked 4th. Funding Impact dimension value is 7, ranked 7th. Because the serial number of papers is sorted in descending order according to the citation, this indicates that the top 1 paper in Scientists Impact has a relatively consistent high impact with each dimension. And considering there are erroneous conclusions in this paper, it may indicate that the wide dissemination of this paper will at the same time cause widespread of erroneous conclusions, which will probably cause relatively serious 'pollution' to science. From the perspective of dimension correlation, in Table 2, the Scientists Impact dimension is highly correlated with Technology Impact and Altmetric Impact, but not with Funding Impact. This result can also be obtained from Fig. 5 and Fig. 6. But whether this rule is universal still needs to be verified on a larger sample set.

Paper 25 ranked top in the Technology Impact dimension, and the dimension value is 18. The result in Scientific Impact dimension is 1, it represents that there are erroneous conclusions that may affect the process of scientific development. The other three dimensions are relatively low. Scientists Impact dimension values is 49, ranked 28th, Altmetric Impact dimension values is 9, ranked 22nd, Funding Impact dimension values is 0. Although paper 25 has the highest value in this dimension, it may not represent the general phenomenon. As can be seen from Figures 4 and 5, except for the paper numbered 25 and published in 1991 with PubMed ID of 2004418, the results of the Technology Impact dimension are generally consistent with those of other dimensions. We further investigated this paper and found that among the 18 patents that cite paper 25, 14 patents share the same inventor. It is very interesting that although this paper was retracted in 1992, one year after its publication year 1991, it was still cited by so many patents in the following years, the earliest priority year of those citing patents varied from 1999 to 2013. The reasons behind this phenomenon, and whether those patents have been affected by retracted papers is worth exploring. From the perspective of dimension correlation, the Technology Impact dimension has a high correlation with the Scientists Impact dimension, a general correlation with the Altmetric Impact dimension, and basically has no correlation with the Funding Impact dimension.

Paper 4 ranked top in the Altmetric Impact dimension, and the dimension value is 460. The result in Scientific Impact dimension is 1, it represents that there are erroneous conclusions that may affect the process of scientific development. Scientists Impact dimension value is 433.5, ranked 2nd, Technology Impact dimension value is 9, ranked 5th, Funding Impact dimension value is 3, ranked 15th. The relationship between the Altmetric Impact dimension and other dimensions of Paper 4 is consistent with the dimension correlation coefficients calculated in Table 2. Altmetric Impact dimension values are highly correlated with Scientists Impact dimension, and are generally correlated with the Technology Impact dimension, but has no correlation with Funding Impact dimension.

Paper 19 ranked top in Funding Impact dimension, and the dimension value is 10. The result in Scientific Impact dimension is 1, it represents that there are erroneous conclusions that may affect the process of scientific development. Scientists Impact dimension value is 71, ranked 22nd, Technology Impact dimension value is 6, ranked 8th, Altmetric Impact dimension value is 158, ranked 2nd. Paper 19 ranks higher in all dimensions except Scientists Impact. However, in terms of correlation, other dimensions are basically not related to Funding Impact dimension. This irrelevance and irregular distribution in the graph may be related to the fact that the amount of funding for each paper has not yet been calculated and the influence of the number of authors has not yet been taken into account, too. More co-authors of a paper may have more funding projects, and it may be more reasonable to calculate the amount of funding for each paper like the method used in Stern et al.(2014). If we calculate the amount of grant for each paper in the case, whether the distribution law and the relevant conclusion_would be consistent with our cognition, that is, paper with higher amount of grant will correspondingly have higher Scientists

Impact or Altmetric Impact, and the person who receives the higher amount of important grant is an influential and important researcher. This assumption is also worthy of further exploration. In the Scientific Impact dimension, the dimension values of paper 2, 9, 22 and 26 are 0, that is to say, these papers do not have wrong conclusions that affect the process of scientific development. The values of these papers in Technology Impact, Funding Impact and Altmetric Impact are generally medium to low, and the ranking distribution in the Scientists Impact dimension is relatively uniform. Whether this phenomenon is rare or common, and whether there is any law behind it, may also be a question worthy of study.

Table 2. Pearson correlation coefficient between dimensions of observation framework

Pearson Correlation Coefficient	*Scientists Impact*	*Technology Impact*	*Altmetric Impact*	*Funding Impact*
Scientists Impact	1	0.408	0.469	0.057
Technology Impact	0.408	1	0.305	-0.032
Altmetric Impact	0.469	0.305	1	0.188
Funding Impact	0.057	-0.032	0.188	1

In order to intuitively observe the situation of various retraction reasons in each dimension, we counted the average values of papers with different retraction reasons in each dimension. The results are shown in Table 3. Since there is only one paper in the Others category, we mainly compare two categories that is Scientific Error and Academic Misconduct. Overall, papers in Academic Misconduct category have a greater impact on the Scientific Impact dimension and Technology Impact dimension than papers in Scientific Error category, whereas Altmetric Impact and Funding Impact are the opposite. But after we try to remove a relatively discrete extreme high value (460 in Paper 4) of the Altmetric Impact dimension in the Academic Misconducts category, the average value of the Altmetric Impact dimension in this category changes from 26.17 to 17.53, that is to say, papers of Academic Misconducts category have more serious impact than those of Scientific Error category in Scientists Impact, Technology Impact and Altmetric Impact. The opposite result of Funding Impact may be due to data selection, or it may be due to the difference of characteristics among different types of retraction reasons, which needs further study.

Table 3. Average Value of Retraction Reasons in Different Dimensions

Retraction Reason	*Scientific Impact*	*Technology Impact*	*Altmetric Impact*	*Funding Impact*
Scientific Errors	123.85	2.4	39.65	3.45
Academic Misconduct	143.32	3.5	26.17	2.89
Others	114.00	0	4	3

Limitations and Future Work

This study attempts to observe the impact of retraction from multiple dimensions, and establishes an observation framework including five dimensions: Scientific Impact, Scientists Impact, Technology Impact, Funding Impact, and Altmetric Impact. Especially, the introduction of Technology Impact dimension has received little attention at present. In this study, the results of the observation framework are displayed and analysed by using bubble diagrams and other charts showing multi-dimensional data. In this way, we could avoid the limitation of integrating different dimensions of data only in one number. However, the sample size for empirical research is not sufficient enough so far, the effectiveness of indicators needs further demonstration, and the data composition of each dimension is relatively simple. In our future study, we will adopt more complete retracted publications' data to further observe the rules under large-sample dataset and verify this observation framework.

In terms of data and indicators, we will conduct sentiment analysis on citation sentences of retracted papers, which represent the attitude of scientists, and add indicators related to citation sentence sentiment analysis of retracted papers to Scientists Impact dimension. As for other dimensions, we will also supplement more detailed and effective data and indicators, such as calculating the amount of funding for each paper. Moreover, we will compare the result before and after retraction, in order to find the impact scope of the retracted publications more clearly, explore its specific impact, and discover truths behind it.

Acknowledgment

We thank Zhesi Shen, Manman Zhu and other members of Center of Scientometrics from National Science Library, Chinese Academy of Sciences for their important suggestions, and Jin Liu from Institute of Information Engineering, Chinese Academy of Sciences for his meaningful support. This work is supported by NSFC(71473236).

Reference

Azoulay, P., Bonatti, A., & Krieger, J. L. (2017). The career effects of scandal: Evidence from scientific retractions. *Research Policy, 46*(9), 1552-1569. doi:10.1016/j.respol.2017.07.003

Azoulay, P., Furman, J. L., Krieger, J. L., & Murray, F. E. (2012). Retractions. *National Bureau of Economic Research Working Paper Series, No. 18499*. doi:10.3386/w18499

Bar-Ilan, J., & Halevi, G. (2017). Post retraction citations in context: a case study. *Scientometrics, 113*(1), 547-565. doi:10.1007/s11192-017-2242-0

Bar-Ilan, J., & Halevi, G. (2018). Temporal characteristics of retracted articles. *Scientometrics, 116*(3), 1771-1783. doi:10.1007/s11192-018-2802-y

Fan, S., & Zhang, Z. (2014). The Distribution of Retracted Articles in Scientific Literature Database. *Journal of the China Society for Scientific and Technical Information, 33*(04), 375-387.

Jin, G. Z., Jones, B., Lu, S. F., & Uzzi, B. (2013). The Reverse Matthew Effect: Catastrophe and Consequence in Scientific Teams. *Social Science Electronic Publishing*.

Luwel, M., van Eck, N. J., & van Leeuwen, T. N. (2018). *The Schön case: Analyzing in-text citations to papers before and after retraction*. Paper presented at the 23rd International Conference on Science and Technology Indicators(STI 2018), Leiden, the Netherlands.

Mongeon, P., & Larivière, V. (2016). Costly collaborations: The impact of scientific fraud on co-authors' careers. *Journal of the Association for Information Science and Technology, 67*(3), 535-542. doi:10.1002/asi.23421

Shuai, X., Rollins, J., Moulinier, I., Custis, T., Edmunds, M., & Schilder, F. (2017). A Multidimensional Investigation of the Effects of Publication Retraction on Scholarly Impact. *Journal of the Association for Information Science and Technology, 68*(9), 2225-2236. doi:10.1002/asi.23826

Stern, A. M., Casadevall, A., Steen, R. G., & Fang, F. C. (2014). Financial costs and personal consequences of research misconduct resulting in retracted publications. *Elife, 3*, e02956. doi:10.7554/eLife.02956

Wager, E., Barbour, V., Yentis, S., & Kleinert, S. (2009). Retractions: Guidance from the Committee on Publication Ethics (COPE). *Maturitas, 64*(4), 201-203. doi:https://doi.org/10.1016/j.maturitas.2009.09.018

Wright, N. D. (1991). *A citation context analysis of retracted scientific articles.* (Ph. D.), University of Maryland College Park, ProQuest Dissertations and Theses database.

Zhang, C., Ding, K., & Liu, Z. (2018). Research Review on Infometric Analysis of Retracted Publications. *Documentation Information and Knowledge*(04), 53-66.

Studying the embeddedness of researchers' careers: Can bibliometric methods help?

Grit Laudel[1]

[1] *grit.laudel@tu-berlin.de*
Department of Sociology, TU Berlin, Fraunhoferstraße 33-36, 10587 Berlin (Germany)

Abstract
Investigations of researchers' careers continue to be shaped by three frames of general career research, namely the neglect of the content of work, a narrow conceptualisation of career success, and a focus on isolated individual careers. The sociology of science and bibliometrics know different. The aim of this paper is to demonstrate that individual-level structural bibliometrics can help us understand the dialectics of ECRs becoming independent while being embedded in knowledge-producing communities and collaborative networks. The paper demonstrates that becoming independent entails a recombination of knowledge acquired in the phase of dependent research (which means keeping cognitive links to one's previous work environments) and its application to new problems (which means establishing an independent research trail). These processes are field-specific.

Introduction

The study of academic careers emerged as an application of general career research to science and higher education. This application carried over three major frames, which increasingly hinder our understanding of academic careers. First, general career research abstracts from the content of work because in most other careers, the content of work is prescribed by organisations and bound to a specific position in the organisation. Second, since the content of work is ignored, career success is reduced to a narrow set of abstract concepts like influence, status, income, or happiness. Third, for the same reason careers are considered as isolated sequences of organisational positions, and their embeddedness in larger work contexts is neglected.

The sociology of science and bibliometrics have shown these assumptions to be wrong, and to demonstrate the embeddedness of researchers and their careers into larger knowledge structures and scientific communities (Velden and Lagoze 2013, Gläser and Laudel 2015a, Cañibano 2017, Aman 2018). The aim of this paper is to demonstrate how bibliometric methods can help in the study of researchers' embeddedness in their work contexts, and the impact of this embeddedness on their careers.

The paper applies individual-level bibliometric methods to the study of the most important thematic change that occurs in the early career phase, namely the emergence of individual research programmes (IRPs). In many research fields, there is a strong expectation of the scientific community that its early career researchers (ECRs) become intellectually independent. Independent researchers formulate research problems and select approaches to solving them on the basis of their own scientific judgement, which is shaped but not overridden by communication and collaboration with peers. For ECRs, becoming independent involves developing an individual research programme that is sufficiently different from those of research groups they had previously worked with as PhD students and postdocs. Previous research has shown that researchers develop their first IRP while conducting dependent research in others' research groups, and that they use the knowledge acquired during this career phase (Laudel and Bielick 2018). Can we find bibliometric traces of this transition to intellectual independence under conditions of embeddedness?

In the last decade, bibliometric studies have paid more attention to the level of individual researchers, often with a focus on external characteristics of publications such as research activity and citation impact (e.g. Glänzel et al. 2018). Only few attempts have been made to bibliometrically measure a researcher's thematic changes over the course of their career

(Hellsten et al. 2007; Gläser and Laudel 2009; Franzoni et al. 2010; Horlings and Gurney 2013; Gläser and Laudel 2015a, Leahey et al. 2008, Van den Besselaar and Sandström 2019). Even fewer studies applied bibliometrics to the systematic study of careers (e.g. Laudel and Gläser 2008; Leahey et al. 2008; Laudel and Bielick 2018).

Theoretical background

Previous studies that attempted to determine an individual's changes of research topics confront us with a confusing variety of terms including research agenda (Horta and Santos 2016), problem area (Horlings and Gurney 2013), research path (Franzoni et al. 2010), and research portfolio (Franzoni and Rossi-Lamastra 2017). These terms are either not defined at all or defined *ad hoc* without any reference to career theory. However, if we consider thematic changes as results of strategic research actions which are part of a researcher's career, they must be accounted for by career theory.

My starting point is a career model that distinguishes three interrelated careers of a researcher (Laudel and Gläser 2008; Gläser and Laudel 2015b). The *cognitive career* consists of thematically connected problem solving processes in which findings from earlier projects serve as input in later projects. These networks of interconnected problem solving processes constitute one or several distinct 'research trails' (Chubin and Connolly 1982), which form a diachronic structure that gradually extends a researcher's knowledge base. The evolution of the content of research has distinct stages and structures, and is closely linked to other career experiences. The *community career* is a series of status positions in the scientific community that are defined by the reputation a researcher has accrued and corresponding role expectations. This includes the status of an apprentice, a colleague, a master and member of the elite. The *organisational career* is a sequence of organisational positions which, through organisational role expectations, are linked to expectations concerning the conduct and content of research and provide opportunities to conduct research (access to salary, infrastructure, and other resources).

In the cognitive strand of their early career phase, researchers are expected to develop their first IRP. An IRP is a researcher's plan for future research that exceeds the scope of a normal project in its thematic breadth and duration. It forms a set of "doable (or fruitful) problems to form a durable identity" (Hackett 2005: 791). An IRP can be characterised by its basis (e.g. a new empirical object or a new experimental system), by the specificity of research problems, methods and objects, by its time horizon (from five years to entire research life), by its thematic breadth (the number of research trails it includes), and by its degree of strategic and technical uncertainty (Laudel and Bielick 2018: 981). Researchers who begin to realise an IRP make a transition from the role of dependent to the role of independent researcher within their community career. In many research fields, this role shift includes the move from dominantly empirical research activities towards more theoretical-conceptual activities (Laudel 2001: 765-766). In their organizational career this transition is reflected by moving from a doctoral and postdoctoral position to a position of independent group leader.

Methods and data

Approach

We construct graphical representations of research trails as a tool that supports the discussion of thematic change in interviews with researchers. Our approach is based on bibliographic coupling, which has been established as one of the best indicators for thematic similarity (e.g. Boyack and Klavans 2010). We collect a researcher's publication oeuvre and construct visual representations of a researcher's cognitive career by identifying and visualising individual research trails. The relative strength of thematic connections was determined using Salton's Cosine (for an extended description of the approach see Gläser and Laudel 2009, 2015).[i]

To support the discussion of thematic change under the time constraints in an interview, we raise the threshold for bibliographic coupling until the main component of the bibliographic coupling network is partitioned into components (disjoint subgraphs that consist of bibliographically coupled publications). This strategy might be less than ideal from a cluster-analysis perspective. However, its purpose is not to produce a perfect thematic delineation[ii] but to trigger narratives about research developments. Still, the overall agreement was remarkably good in many specialities we studied over the last 15 years (particularly fields within experimental physics and biology). Interviewees rarely corrected our representations by linking or splitting clusters. Researchers in the experimental sciences usually had cognitive careers consisting of one, two or up to four parallel research trails.

Determining an ECR's transition to independence requires studying the embeddedness of their research into the research programmes of their former group leaders. This can be achieved by analysing coupled research trails (Gläser and Laudel 2015: 324-325). An ECR's transition to independence should be represented by the loss or reduction of their research trails' strong connections to their former group leader's research.

The role shift from dominantly empirical work toward conceptual work described above is reflected in the co-authorship order. The first author is usually the researcher who carried out the empirical work (typically the PhD student or postdoc) and the last author is the researcher who did mainly conceptual work (typically the group leader). Middle authorship is usually given for other collaborative contributions (Stokes and Hartley 1989; Laudel 2002). Thus, when ECR change their role from dependent researcher to independent group leader in their organisational career, a move from dominantly first authorship towards dominantly last authorship can be expected.

Case selection

I draw on empirical data from three empirical studies on researchers' early career phase. These studies focus on how ECRs realise thematic changes (Laudel 2017), develop IRPs (Laudel and Bielick 2018) and – in an ongoing project - how they realise their first IRPs.

I included 15 cases from three fields: experimental Atomic and Molecular Optics, plant biology and theoretical chemistry. Within these fields I selected ECRs who – according to my interview data – had started their own research group and have already been working on their first IRP for some time (four to ten years). [iii] This increased the likelihood that their research is bibliometrically visible in form of publications. For each ECR I searched for the names of their group leaders from the groups where they conducted dependent research as PhD students or as postdocs.

Methods and data

For each ECR and their senior group leaders I downloaded the publications in the Web of Science (core collection). Only publications of the document type "article" were included.

Some senior group leaders had a very extensive publication oeuvre which led to a confusing, hard to handle network graphics. For those, we included only publications on which they appeared as last author. The underlying assumption – which we tested - is that this depicts their core work and excludes mere collaborative contributions to others' work (Laudel 2002).

Three analyses were conducted for each ECR:

1. The development of the authorship order of the ECR's publications over time was analysed to answer the question as to whether and when the ECR gained last authorship positions.
2. The ECR's individual research trails were constructed in order to check the extent to which reported thematic changes or continuities are visible in these research trails.
3. The changing embeddedness of the ECR's research was analysed by constructing combined research trails for the ECR and each of their senior group leaders. This was achieved by creating a joined dataset of the publications of two researchers and identifying research trails in this dataset.

I experimented with varying thresholds of the Salton value in order to find stable clusters. To avoid artefacts, three criteria were applied in the selection of thresholds. First, most publications of a researcher should be bibliographically coupled rather than isolated. Second, isolated publications (publications that were not bibliographically coupled when a particular threshold was applied) were checked for being thematic outliers (whose isolation was often caused by making collaborative contributions to others' research) or being so recent that they might indicate the beginning of a new cluster. Third, the cluster solutions of the coupled research trails were checked for plausibility by triangulating them with accounts provided by the interviewed ECRs.
The comparative case studies were based on semi-structured interviews, which lasted on average 90 minutes and consisted of two main parts. In the first part, the interviewee's research and cognitive career were discussed using network representations of the interviewee's research trail. The development of the interviewee's research since the PhD project was explored with a focus on thematic changes and their reasons. The second part of the interview focused on transitions in the organisational career. Data from the semi-structured interviews with the selected ECRs were available in form of the interview transcripts and of a structured information base, derived previously from qualitative content analysis of the interviews. For reasons of confidentiality, all data are provided in anonymised form.

Results

History of co-authorship order indicating transition to independence

The concept of first and last authorship can be applied to all cases because in all three fields the dominant research practice is to work in groups consisting of PhD students and postdocs, led by a group leader. The typical organisational career of the ECR in the investigated fields consists of a first phase of dependent positions (a PhD position, followed by one or more postdoctoral positions) and a second phase of an independent group leader position.
For all 15 ECRs this pattern has been confirmed: researchers moved from predominantly first (and middle) authorships to last authorships after obtaining an independent group leader position (Figure 1).

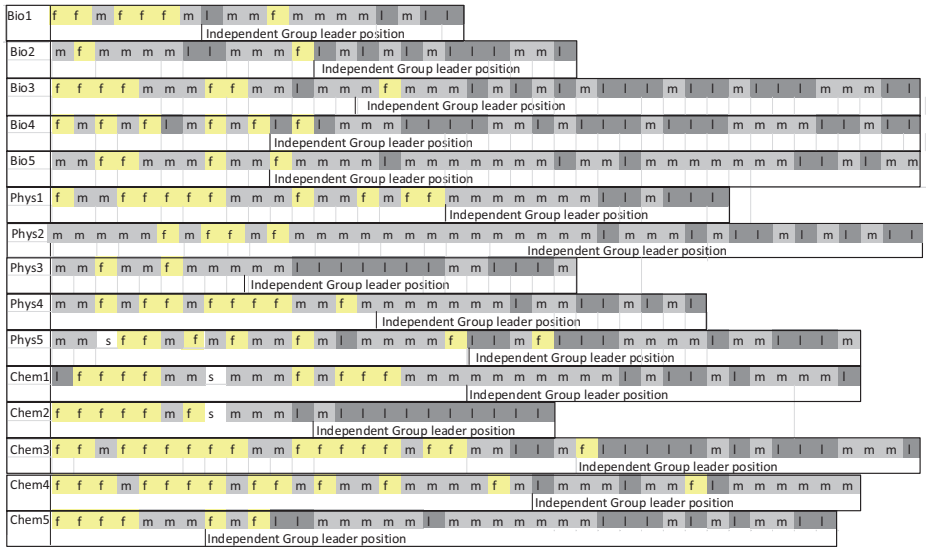


Figure 1. Development of co-authorship order during an ECR's career (up to 40 publications per author, f: first author, m: middle author, l: last author, s: single author)

Researchers kept their middle authorship position when they collaborated with other research groups. This occurs frequently, particularly in areas such as molecular biology where many specialised research methods exist and collaborative contributions to others' work is common. In theoretical chemistry middle authorship also occurred when researchers conducted service collaborations in form of calculations for experimentalists. In experimental AMO physics middle authorships occurred when the ECR's collaborated with theoreticians.

Coupled research trails indicating embeddedness, recombination of knowledge and transition to independence

The analysis of co-authorship patters in the three fields showed that all ECRs founded their own research groups, and that some of them remained embedded in collaborative relationships with their former group leaders. This raises two questions. First, did the ECRs who became formally independent also become intellectually independent from their former research groups? Second, how are the ECRs' IRPs informed by their work as dependent researchers in their former research groups? The following two examples illustrate how this can be investigated bibliometrically.

Our first example is an AMO physicist (Physicist 2). This ECR had conducted his PhD in atomic physics about a problem concerning interacting ultracold atoms. His subsequent postdoctoral research he described as a rather strong change from the subfield of pure atomic physics towards the subfield of quantum optics. During his first postdoc he developed an IRP which continues work in quantum optics, but combines it with atomic physics, i.e. with techniques from his PhD and with techniques he learned during a short second postdoc.

His research trail (Figure 2) shows all these developments: the PhD work (red cluster), the postdoctoral work (green cluster) and the work on his IRP (yellow cluster). In this most recent cluster he also became last author on some publications.

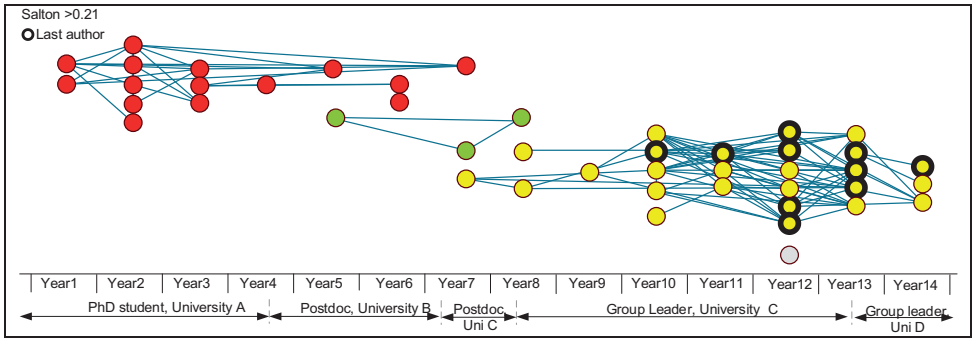


Figure 2. Research Trail of AMO Physicist 2

This is the traditional 'isolationist' framing of ECRs' careers, which so far could only be overcome by interviews tracing intellectual influences (Laudel 2017, Laudel and Bielick 2018). We can now try to link these interview accounts to further bibliometric analyses. The bibliographic coupling of the ECR's publications with the publications of his former PhD group leader (Figure 3) shows the group leaders publication on the top (red area), the ECR's publications on the bottom (yellow area) and joint publications in the middle (overlapping area). With his PhD topic, the ECR obviously joined an area that the group leader has already been working on. He left this research area after he finished his PhD (the four publications that were published during his postdoc time still belong to his PhD topic). Figure 3 shows a clear separation of the PhD group leader's research program and the ECR's IRP (yellow cluster).

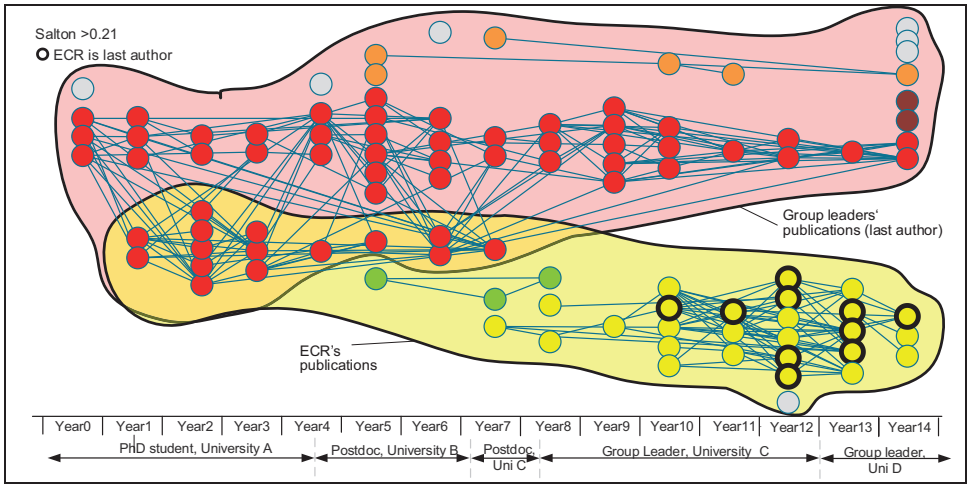


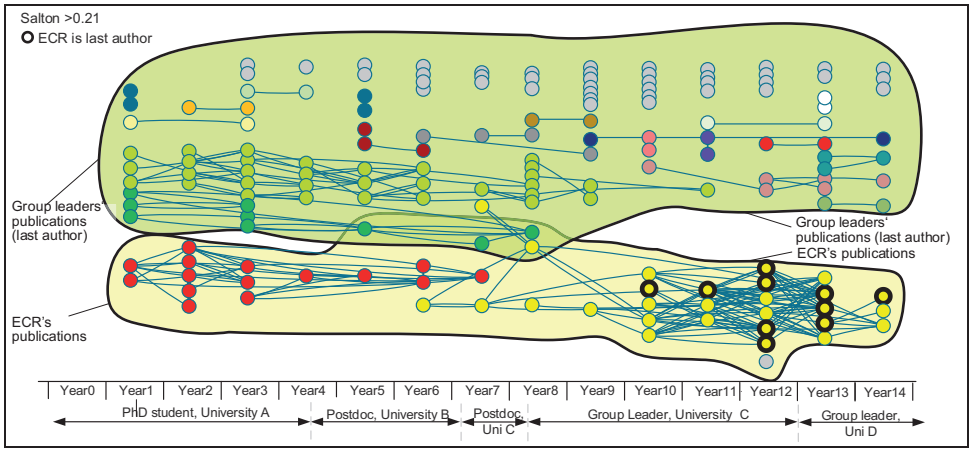
Figure 3. Combined research trails of AMO Physicist 2 and his PhD group leader

The ECR could start his IRP already during his first postdoc. His postdoctoral group leader became very interested in the ECR's research idea, contributed conceptual ideas to it and gave the ECR the opportunity to build an experimental setup during this time. This setup stayed in the postdoctoral group, and the ECR later built a second one when he became independent group leader. Both, the former postdoctoral group and the ECR's group conduct similar experiments based on the same experimental system in principle, but adding slight modifications. However – and this is common in AMO physics – it allowed to pursue different sets of problems to avoid direct competition.

> Yes, in that sense we are direct competitors. Still, our contact is very friendly and very peaceful. [...] That's also not unusual in our area, I would say, that people continue with something very similar to what they have learned in their doctoral or postdoctoral studies. They become independent with experiments where they continue something because the setups are so complicated, and you have to start somewhere. [...] We have clearly chosen a topic that is orthogonal to what was done there [in the postdoctoral group]. So this [topic] was a clear continuation of this work, but in the [postdoc] group they have clearly moved into another direction. We have said very consciously, we do not do exactly the same thing they do. (Physicist 2)

Are these similarities and differences visible if we combine both groups' research trails?

Figure 4 shows that the ECR's subsequent publications after his postdoc are still bibliographically connected to the publications of his postdoctoral time (yellow cluster). His postdoctoral group leader has no publications anymore that are connected to this cluster. The bibliographic coupling suggests that both groups pursue different IRPs. If we lower the Salton value (>0.18), then the ECR's publication from his career phase as a postdoc and as a group leader become more connected to his postdoc group leader's main publication cluster, while his PhD publications remain a separate cluster. It is clearly visible how the ECR's IRP is rooted in the postdoc group leader's IRP, yet has separated by now.



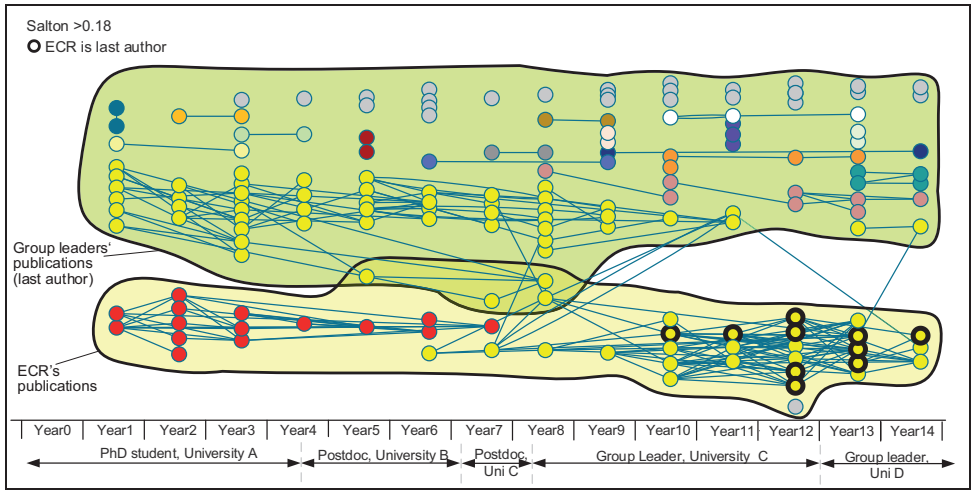


Figure 4. Combined research trails of AMO Physicist 2 and the group leader of his first postdoc (with two different Salton values)

The second example shows the research trail of a theoretical chemist (Figure 5). During his PhD this chemist developed a new approach for studying the electronic structure of a class of molecules, which was suggested by his group leader. Although he could only partly solve the problem, two major publications came out of this work (small red cluster). He then moved on to do a postdoc in order to learn a completely different set of methods and applying these methods to experimental chemical data (the start of the green cluster). In this period he also learned yet another computational method and developed it further for computations of electronic structures. During his postdoc he still collaborated with his PhD group leader who suggested the application of methods similar to those he used in his PhD work in a new approach to the problem he had worked on during his PhD. In his IRP, the ECR combined methods from his PhD and from his postdoc. It consists of two interlinked research trails, depicted in the publications of the green cluster and the yellow cluster. These two research trails are different sets of combinations of the methods he learned and developed previously. On the publications of these two trails he became last author.

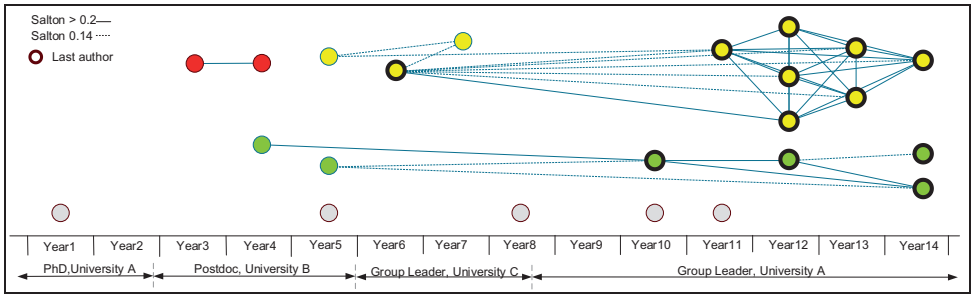


Figure 5. Research Trail of Theoretical Chemist 2

The research trail depicted in the yellow cluster is closely linked to his PhD group leader's publications in form of a large shared cluster (Figure 6). Although they ceased to publish together, their publications are closely connected through shared references. Both try to solve similar problems but using different techniques for that. This difference in the applied methods is not visible in the bibliometric visualisation of their cognitive careers, even after considerably increasing the threshold (the Salton value) for the cluster separation.

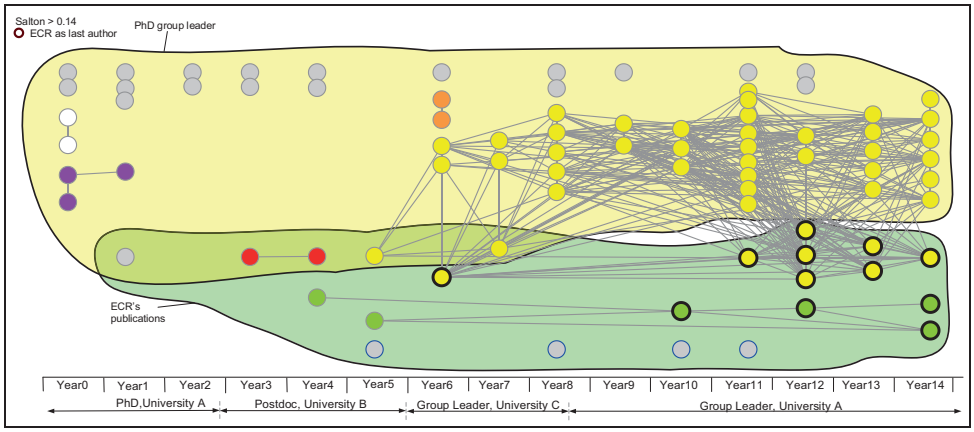


Figure 6. Combined research trails of Theoretical Chemist 2 and his PhD group leader

The combined research trails of chemist 2 and the leader of his postdoctoral group shows that the ECR's postdoctoral work was not connected with the group leader's main area of research (blue cluster in Figure 7). Two publications co-authored with the group leader started the green research trail which the interviewee describes in the following quote:

> And so that was a new problem that needed to be solved, which I hadn't even thought would ever be a problem […]. So we developed some ideas behind that [referring to publication], to try and at least get a handle on that. If we can routinely locate these things, we can understand them a bit. And that started, I suppose, one of the directions I have been carrying on ever since then.

As we can see, this research is not pursued by the group leader of the postdoctoral group. This is not surprising because the research problem was defined by Chemist 2 rather than by his group leader:

> I came up with an idea, I did it, and I put a paper on his desk and said: "How is that?" And he said: "Yep, submit it." That was usually what happened there. So it was just a really nice environment for me to think about problems and talk to people and get ideas but I didn't feel [my group leader] needed to give me any ideas. So that's that strand.

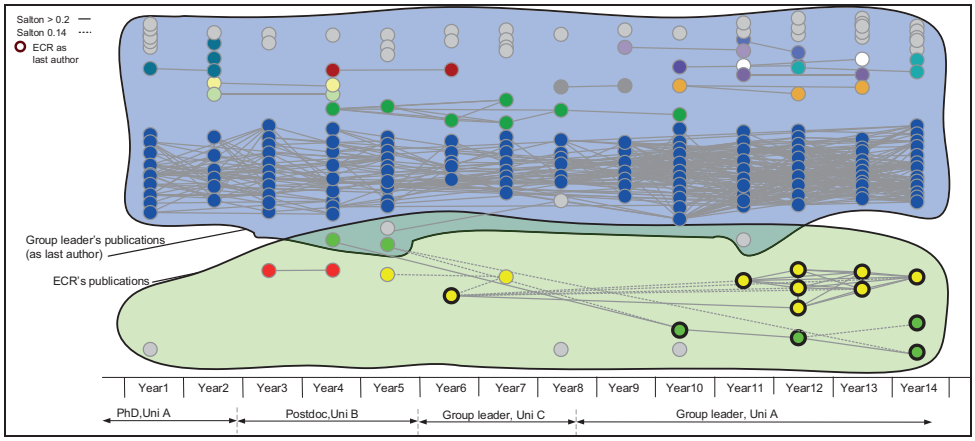


Figure 7. Combined research trails of Theoretical Chemist 2 and his postdoc group leader

Finally, an overview of all analysed cases is provided (Table 1). In the interviews, all 15 ECR explained how their IRP differed from those of their respective group leaders in terms of research problems, objects or methods. At the same time, there are cases in all three fields where the epistemic differentiation is not visible in the bibliographic coupling, i.e. no clusters that are separate from those of former group leaders become visible.

Table 1. Relationship between ECR's and their group leader's research, bibliometrically constructed

Case	*Time of separation from PhD group*	*Time of separation from postdoc group*
Bio1	during postdoc	no separation
Bio2	no separation	while group leader
Bio3	while group leader	while group leader
Bio4 trail 1	n.a. (professor retired)	no separation
trail 2	n.a. (professor retired)	while group leader
Bio5	while group leader	while group leader
Phys1	while group leader	while group leader
Phys2	during postdoc	during postdoc
Phys3	while group leader	while group leader
Phys4	no separation	no separation
Phys5	during postdoc	during postdoc
Chem1	no separation	no separation
Chem2 trail 1	no separation	no separation
trail 2	during postdoc	during postdoc
Chem3 trail 1	during postdoc	no separation
trail 2	during postdoc	no separation
Chem4 trail 1	while group leader	while group leader
trail 2	while group leader	while group leader
Chem5 trail 1	no separation	no separation
trail 2	no separation	while group leader
trail 3	while group leader	while group leader

Discussion

I demonstrated how bibliometric methods can be used to study the embeddedness of academic careers. Co-authorships and coupled research trails show the dialectics of becoming independent and staying connected. All ECRs became last authors on some of their publications. The two examples discussed in more detail reveal that last authorship is connected to research on one's own IRP.

The embeddedness of the ECR's research into their scientific community's research became apparent when we looked at the process of their IRP development. The two ECRs whose cognitive careers I discussed in detail acquired new knowledge and combined it with knowledge from their PhD phase, which sparked ideas for their IRPs. This process was similar in the other cases (Laudel and Bielick 2018).

The bibliometric reconstruction of the ECR's cognitive careers led to clusters in a bibliographic coupling network that matched the descriptions of IRP emergence in the interviews. If the researchers had implemented their IRPs, this became visible in form of new thematic clusters. Becoming independent was reflected in research clusters which started in another group but then became separated. In other cases, the cluster representing the ECR's IRP remained coupled to a cluster of their former group leaders despite stated epistemic differences between the ECR's and the former group leader's research. this does not mean that we either have to disregard the interviewees' accounts or must consider the bibliometric approach unsuitable. It is more likely that the ECRs' IRPs differ from the research of the ECRs' former group leaders with regard to theory *or* method *or* object but not in all three, and that the remaining similarities lead to strong bibliographic coupling. The remaining strong bibliographic links between a group leader's thematic clusters and the ECR's cluster(s) are also an indicator for the embeddedness of the ECR's work. Such links remained relatively frequently in the research trails of the theoretical chemists, which may be an indicator for field-specific interdependencies (Gläser et al. 2018).

There are two further limitations of the bibliometric approach. The first limitation is the time lag. In order to see clusters of publications that represent topics of an ECR's research, the ECR must have worked for several years on this topic and must have produced several publications. In some fields like experimental AMO physics this may take more than six years.

The second limitation is that our current approach requires setting an arbitrary threshold for the construction of components in the bibliographic coupling network. Although my experiments with different thresholds revealed a certain stability of the number of larger clusters, the threshold is more difficult to set if only small clusters of few publications emerge. Using this bibliometric approach as a stand-alone method without any other information about a group's research cannot be recommended.

Conclusions

Sociological research on academic careers is still limited by an 'isolationist' approach that neglects the embeddedness of researchers and their work into the research of their scientific communities, which strongly influences their cognitive, community and organisational careers. Bibliometrics can help to overcome this neglect by providing methods to study the embeddedness of a researcher's cognitive career. An important contribution of this perspective on embedded independence is a better understanding of individual-level and group-level thematic change.

The study of embedded research trails touches upon a major theme of innovation research, namely the emergence of innovations from the recombination of knowledge. However, the investigation presented here also demonstrates that at least in research, mere recombination is not enough. Although all IRPs recombined knowledge the ECRs acquired in their previous research groups, they also had an original element by applying the recombined knowledge to a new subject.

The utilisation of bibliometric methods presented in this paper started from a theoretical concept of careers, cognitive careers and thematic change. There may be other approaches to capturing the cognitive career of a researcher, and there are certainly more advanced methods for clustering a researcher's publications. However, any use of bibliometric methods for studying thematic change requires clarifying and operationalising the underlying theoretical concepts as well as validating the results, i.e. reconstructing the content of research with qualitative methods.

Independence is a phenomenon that is not quantifiable. From this follows that the search for an algorithm that can 'measure' independence regardless of circumstance (e.g. field or career stage) is futile. Reducing the complex phenomenon of independence to a numerical indicator, composite or otherwise, that shall replace evaluations based on peer review (van den Besselaar and Sandström 2019: 17) is not possible.

Although all investigated ECR conducted (independent) research that was different from their former group leaders' research, these differences were not always depicted in the bibliometric representations of research trails. Future research should explore how research programmes of individuals or groups differ in terms of their methodological, theoretical or object-based interdependence, and how these differences are reflected in bibliometric structures.

This approach can also be used to study the interaction between embeddedness and mobility (for first attempts see Aman 2018).

Finally, this research raises an interesting question about evaluations: If the research of individuals and the groups in which they worked is so strongly embedded, it may well be that even research groups are difficult to evaluate in isolation.

Notes

[i] Previous attempts to determine individual researcher's topics used broad keyword descriptors of journals (Leahey, Crockett, and Hunter 2008) which is far too coarse and unreliable for our purposes. Others applied more sophisticated clustering methods and combined them with modularity-maximising algorithms (Franzoni et al. 2010, Horlings and Gurney 2013, Van den Besselaar and Sandström 2019), which broke up the research into an unrealistic large number of topics per researcher. The validity of these approaches has not been investigated.

[ii] This is impossible because research topics do not exist independently from the perception of the researcher who works on them.

[iii] I did not include "unsuccessful" cases, i.e. ECR who had not started to work independently on their IRP because they had only co-authored publications with their group leaders or ceased to publish when leaving academia.

References

Aman, Valeria (2018). *Transfer of formal knowledge through international scientific mobility - introduction of a network-based bibliometric method.* 23rd International Conference on Science and Technology Indicators (STI 2018), September 12-14, Leiden, The Netherlands, 545-552.

Boyack, Kevin W., and Richard Klavans (2010). Co-citation analysis, bibliographic coupling, and direct citation: Which citation approach represents the research front most accurately? *Journal of the American Society for Information Science and Technology,* 61, 2389-2404.

Cañibano, Carolina (2017). Scientific mobility and economic assumptions: From the allocation of scientists to the socioeconomics of network transformation. *Science as Culture* 26: 505-519.

Chubin, Daryl E., and Terence Connolly, 1982. Research Trails and Science Policies. Norbert Elias, Herminio Martins and Richard Whitley (eds.), *Scientific Establishments and Hierarchies*. (pp. 293-311). Dordrecht: Reidel.

Franzoni, Chiara, and Cristina Rossi-Lamastra, 2017. Academic tenure, risk-taking and the diversification of scientific research. *Industry and Innovation,* 24, 691-712.

Franzoni, Chiara, Christopher Simpkins, Baoli Li, and Ashwin Ram (2010). Using content analysis to investigate the research paths chosen by scientists over time. *Scientometrics,* 83, 321-335.

Glänzel, Wolfgang, Bart Thijs, Lin Zhang, and Gunnar Sivertsen (2018). *Developing a model for individual-level measurement of research activity and citation impact*. 23rd International Conference on Science and Technology Indicators (STI 2018), September 12-14, Leiden, The Netherlands.

Gläser, Jochen, and Grit Laudel, (2009). Identifying individual research trails. Birger Larsen and Jacqueline Leta (eds.), *Proceedings of the 12th International Society for Scientometrics and Informetrics (ISSI'09)*. Rio de Janeiro, 841-845.

Gläser, Jochen, and Grit Laudel (2015a). A Bibliometric Reconstruction of Research Trails for Qualitative Investigations of Scientific Innovations. *Historical Social Research / Historische Sozialforschung,* 40, 299-330.

Gläser, Jochen, and Grit Laudel (2015b). *The Three Careers of an Academic*. Discussion Paper 35/2015. Berlin: TU Berlin, Center for Technology and Society, https://www.tu-berlin.de/fileadmin/f27-/PDFs/Discussion_Papers/35_2015discussion_paper_Nr_35_Glaeser_Laudel.pdf.

Gläser, Jochen, Grit Laudel, Christopher Grieser, and Uli Meyer (2018). *Scientific fields as epistemic regimes: new opportunities for comparative science studies*. The Technical University Technology Studies Working Papers, TUTS-WP-3-2018, Berlin: Technical University Berlin, https://www.ts.tu-berlin.de/fileadmin/i62_tstypo3/TUTS-WP-3-2018_Scientific_Regimes.pdf.

Hackett, Edward J. (2005). Essential Tensions: Identity, Control, and Risk in Research. *Social Studies of Science,* 35, 787-826.

Hellsten, Iina, Renaud Lambiotte, Andrea Scharnhorst, and Marcel Ausloos (2007). Self-citations, co-authorships and keywords: A new approach to scientists' field mobility? *Scientometrics,* 72, 469-486.

Horlings, Edwin, and Thomas Gurney (2013). Search strategies along the academic lifecycle. *Scientometrics,* 94, 1137-1160.

Horta, Hugo, and João M Santos (2016). An instrument to measure individuals' research agenda setting: the multi-dimensional research agendas inventory. *Scientometrics,* 108, 1243-1265.

Laudel, Grit (2001). Collaboration, creativity and rewards: why and how scientists collaborate. *International Journal of Technology Management,* 22, 762-781.

Laudel, Grit (2002). What do we measure by co-authorships? *Research Evaluation,* 11, 3-15.

Laudel, Grit, and Jana Bielick (2018). The emergence of individual research programmes in the early career phase of academics. *Science, Technology, & Human Values,* 43, 972-1010.

Laudel, Grit, and Jochen Gläser (2008). From apprentice to colleague: the metamorphosis of Early Career Researchers. *Higher Education,* 55, 387-406.

Leahey, Erin, Jason Lee Crockett, and Laura Ann Hunter (2008). Gendered Academic Careers: Specializing for Success? *Social Forces,* 86, 1273-1309.

Stokes, T. D., and J. A. Hartley (1989). Coauthorship, Social Structure and Influence within Specialties. *Social Studies of Science,* 19, 101-125.

Van den Besselaar, Peter, and Ulf Sandström (2019). Measuring researcher independence using bibliometric data: A proposal for a new performance indicator. *PloS one,* 14, e0202712.

Velden, Theresa, and Carl Lagoze (2013). The Extraction of Community Structures from Publication Networks to Support Ethnographic Observations of Field Differences in Scientific Communication. *Journal of the American Society for Information Science and Technology,* 64, 2405-2427.

Community Detection Using Citation Relations and Textual Similarities in a Large Set of PubMed Publications

Per Ahlgren[1], Yunwei Chen[2], Cristian Colliander[3], and Nees Jan van Eck[4]

[1] *per.ahlgren@uadm.uu.se*
Uppsala University, Department of Statistics, Ekonomikum (plan 3), Kyrkogårdsg 10, Uppsala (Sweden)

[2] *chenyw@clas.ac.cn*
Scientometrics & Evaluation Research Center (SERC), Chengdu Library and Information Center of Chinese Academy of Sciences, Chengdu, 610041 (China)

[3] *cristian.colliander@umu.se*
Department of Sociology, Inforsk, Umeå University, Umeå (Sweden)
University Library, Umeå University, Umeå (Sweden)

[4] *ecknjpvan@cwts.leidenuniv.nl*
Centre for Science and Technology Studies, Leiden University (The Netherlands)

Abstract

In this contribution, the effects of enhancing direct citations, with respect to publication-publication relatedness measurement, by indirect citation relations (bibliographic coupling and co-citation) and text relations on clustering accuracy are analyzed. In total, we investigate six approaches. In one of these, direct citations are enhanced by both bibliographic coupling and co-citation, whereas text relations are used to enhance direct citations in another approach. In addition to an approach based on direct citations only, we include in the study, for comparison reasons, each approach that is involved in the enhancement of direct citations. For the evaluation of the approaches, we use a methodology proposed by earlier research. However, the used evaluation criterion is based on MeSH, arguable the most sophisticated item-level classification scheme available. The results show that the co-citation approach has the worst performance, and that the direct citations approach is outperformed by the other four investigated approaches. An approach in which direct citations are enhanced by the BM25 textual relatedness measure has the best performance, followed by the approach that combines direct citations with bibliographic coupling and co-citation. The latter performs slightly better than the bibliographic coupling approach, which in turn has a better performance than the BM25 approach.

Introduction

Community detection in citation networks, which is the topic of this paper, can be performed in order to reveal both the obvious and the more subtle interrelations between subfields of science, as well as the growth and decline of such subfields (Chen & Redner, 2010). In the context of networks, communities are clusters of closely connected nodes within a network. Communities of this kind are found, not only in citation networks, but also in many other networks, like biological networks, the World Wide Web, social networks and collaboration works (Girvan & Newman, 2002).

Citation networks originate from the relationships between citing and cited publications. In these networks, community structure can often be observed, since publications dealing with a given topic tend to cite similar publications with respect to topic. Communities in a citation network thereby contain similar publications regarding a single topic or a set of related topics. For a given field, community detection in a citation network can be used for uncovering related publications. The detected subfields might then be useful for researchers and policy makers, since they indicate the whole pattern of the field at a glance.

Although several studies on community detection in citation networks have been performed in recent years, we have not found many such studies that discriminate, based on some notion of

importance, between direct citations. However, Small (1997) explored idea of combining direct citation information with indirect citation information. Persson (2010) used weighted direct citations, where the citations were weighted by shared references and co-citations in order to decompose a citation network. Persson investigated the field of library and information science and obtained meaningful subfields by removing direct citations with weights below a certain threshold and by removal of less frequently cited publications. The study by Fujita et al. (2014) constitutes as another example of a study using weighted direct citations. Different types of weighted citation networks were studied with regard to detection of emerging research fields, where the weights were based on, for instance, reference lists and keyword similarity. Chen et al. (2013) proposed a community discovery algorithm to uncover semantic communities in a citation semantic link network. In that study, direct citations were weighted on the basis of common keywords. A fifth example of a study that discriminates between direct citations is the work by Chen et al. (2017), which inspired us to perform the study of this paper. These authors used two publication data sets and modularity-based clustering of publications, and compared clustering solutions obtained on the basis of four approaches, where the main difference between these approaches is how the relatedness of two publications is defined. One of the approaches is based on direct citations, whereas the other three weight the direct citations in three different ways. All the latter three approaches use textual similarities as weights, and two of them take term position information into account.

One can distinguish between two types of methods used for citation network community detection. One type consists of methods based only on the topological structure of the network, i.e. the arrangement of publications (nodes) and direct citations (links) (e.g. Boyack & Klavans, 2014; Chen & Redner, 2010; Kajikawa et al., 2008; Kusumastuti et al., 2016; Ruiz-Castillo & Waltman, 2015; Sjögårde & Ahlgren, 2018; Subelj et al., 2016; Waltman & Van Eck, 2012; Yudhoatmojo & Samuar, 2017), whereas the other type consists of methods that also use publication content, represented by text. To take both topological structure and content into account in an analysis of citation networks might be fruitful. This has been done, as we have seen, in community detection analyses (Chen et al., 2013; Fujita et al., 2014; Chen et al., 2017). However, it has been done also in studies not involving community detection. Cohn and Hofmann (2001) described a joint probabilistic model for modelling the contents and inter-connectivity of publication collections such as sets of research publications, while Hamedani et al. (2016) presented a novel method called SimCC that considers both citations and content in the calculation of publication-publication similarity.

Even if the last two papers referred to in the preceding paragraph did not involve community detection in networks, they provide ideas that can be used for community detection in such networks. Indeed, in this study we use both topological structure and content information in citation networks to detect communities. We build on the earlier work by Chen et al. (2017) on weighting of citation relations, as well as on the work by Waltman et al. (2017, 2019) on a principled methodology for evaluating the accuracy of clustering solutions using different relatedness measures. In this study, the effects of enhancing direct citations, with respect to publication-publication relatedness measurement, by indirect citation relations and text relations on clustering accuracy are analyzed. In total, we investigate six approaches. In one of these, direct citations are enhanced by both bibliographic coupling and co-citation, whereas text relations are used to enhance direct citations in another approach. In addition to an approach based on direct citations only, we include in the study, for comparison reasons, each approach that is involved in the enhancement of direct citations. Compared to the study by Chen et al. (2017), a considerably larger publication set is used in our study, as well as a more sophisticated evaluation methodology, in which an external subject classification scheme, Medical Subject

Headings (MeSH), is used. MeSH is arguable the most sophisticated item-level classification scheme available. Moreover, in contrast to the earlier work, we use a different approach regarding the combination of direct citations and text relations. Compared to Waltman et al. (2017, 2019), these authors did not evaluate hybrid relatedness approaches (approaches combining citation and text relations). Further, citation-only approaches were only compared to other such approaches in their analysis, and the same was the case for text-only approaches. In our study, however, comparisons across such approach groups are made, due to the use of MeSH as an independent evaluation criterion.

The remainder of the paper is organized as follows. In the next section, we deal with data and methods, whereas the results of the study are reported in the third section. In the final section, we provide a discussion as well as conclusions.

Data and methods

Since direct citations are used in the study, we needed a sufficiently long publication period. We decided to use a five-year period, namely 2013-2017. Initially, a set of 4,260,452 MEDLINE–the largest subset of PubMed–publications were retrieved from PubMed, where the query included a reference to the publication period. The following query was used: *MEDLINE[SB] AND ("2013/01/01"[PDat] : "2017/12/31"[PDat])*. From the initially retrieved set, we filtered out those publications with a print year in the interval 2013-2017, which yielded a set of 4,191,763 publications. Since PubMed does not contain citation relations between publications, we also use Web of Science (WoS) data. The next step was then to match, using PMID data, each publication in this set of publications to publications included in the in-house version of WoS database available at the Centre for Science and Technology Studies (CWTS) at Leiden University, which yielded a set of 3,577,358 publications. From this latter set, we selected each publication *d* such that *d* satisfies each of the following four conditions:

1. *d* has a WoS publication year in the period 2013-2017.
2. *d* is of WoS document type *Article* or *Review*.
3. *d* has both an abstract and a title with respect to its WoS record.
4. *d* has a citation relation to at least one publication *d'* such that *d'* satisfies points 1-3 in this list.

2,941,119 publications satisfied all four conditions. However, 10 of these publications were removed, since they are not indexed with MeSH descriptors. Such descriptors are needed by our evaluation methodology (see the subsection "Evaluation of approach performance" below). Our final publication set, say P_{MEDLINE}, then consists of 2,941,109 publications.

Investigated approaches

As stated above, we compare six approaches to publication community detection in this study. The main difference between the approaches is how the relatedness of two publications is defined. Four of the approaches–DC (direct citation), BC (bibliographic coupling), CC (co-citation) and DC-BC-CC–use only citation relations. For the remaining two approaches, BM25 and DC-BM25, BM25 uses only text relations, whereas DC-BM25 combines direct citations with text relations. We now describe the six approaches in more detail.

DC. In DC, the relatedness of two publications *i* and *j*, r_{ij}^{DC}, is 1 if there is a direct citation from *i* to *j* or such a relation from *j* to *i*, otherwise the relatedness is 0.

BC. Here, the relatedness of *i* and *j*, r_{ij}^{BC}, is defined as the number of shared cited references in *i* and *j*, where only cited references pointing to publications covered by the CWTS in-house version of WoS are taken into account.

CC. The relatedness of *i* and *j*, r_{ij}^{CC}, is defined as the number of publications that cite both *i* and *j*.

BM25. The first step in this approach is to identify terms in the titles and abstracts of the publications in P_{MEDLINE}. Here a *term* is defined as a noun phrase: a sequence *s* of words of length *n* ($n \geq 1$) such that (a) each word in *s* is either a noun or an adjective, and (b) *s* ends with a noun. The part-of-speech tagging algorithm provided by the Apache OpenNLP 1.5.2 library is used to identify the nouns and adjectives. Plural and singular noun phrases are regarded as the same term, and shorter terms appearing in longer terms are not counted.

The BM25 approach involves the BM25 measure, a well-known query-publication similarity measure in information retrieval research (Sparck Jones, Walker, & Robertson, 2000a, 2000b) and, according to experimental results obtained by Boyack et al. (2011), one of the most accurate text-based measures for clustering publications. Let *N* be the number of publications under consideration (in our case, *N* is equal to $|P_{\mathrm{MEDLINE}}| = 2{,}941{,}109$) and *m* the number of unique terms occurring in the *N* publications. Let o_{il} be the number occurrences of term *l* in publication *i*, and n_l the number of publications in which term *l* occurs. Further, $I(o_{il} > 0) = 1$ if $o_{il} > 0$ and 0 otherwise. The relatedness of *i* and *j*, r_{ij}^{BM25}, is then defined as

$$r_{ij}^{\mathrm{BM25}} = \sum_{l=1}^{m} I(o_{il} > 0)\ \mathrm{IDF}_l \frac{o_{jl}(k_1 + 1)}{o_{jl} + k_1\left(1 - b + b\dfrac{d_j}{\overline{d}}\right)} \tag{1}$$

where

$$\mathrm{IDF}_l = \log \frac{N - n_l + 0.5}{n_l + 0.5} \tag{2}$$

and

$$d_j = \sum_{p=1}^{m} o_{jp};\ \overline{d} = \frac{1}{N}\sum_{q=1}^{N}\sum_{p=1}^{m} o_{qp} \tag{3}$$

IDF_l is the inverse document frequency of term *l*, d_j the length of publication *j*, and $\overline{d}$ the mean length of the *N* publications. k_1 and *b* are parameters with respect to term frequency saturation and publication length normalization, respectively. For the values of these, we followed Boyack et al. (2011) and Waltman et al. (2017, 2019), and thereby used 2 and 0.75 for k_1 and *b*, respectively. Note that it is possible that $r_{ij}^{\mathrm{BM25}} \neq r_{ji}^{\mathrm{BM25}}$, i.e. the BM25 measure is not symmetrical. It follows from Eq. (1) that $r_{ij}^{\mathrm{BM25}} > 0$ if and only if there is at least one term occurring in both *i* and *j*.

DC-BC-CC. In this approach, direct citations are enhanced by the citation relations corresponding to the approaches BC and CC. We define relatedness of *i* and *j*, $r_{ij}^{\text{DC-BC-CC}}$, as

$$r_{ij}^{\text{DC-BC-CC}} = \alpha r_{ij}^{\text{DC}} + r_{ij}^{\text{BC}} + r_{ij}^{\text{CC}} \tag{4}$$

where α is a weight of direct citations relative to BC and CC. With this weight, one has the possibility to boost direct citations, which might be considered as stronger signals of the relatedness of two publications compared to a bibliographic coupling or a co-citation relation (Waltman & van Eck, 2012). In our analysis, we use 1 and 5 as values of α, in agreement with Waltman et al. (2017, 2019). Note, in contrast to DC, that the relatedness value of *i* and *j* in DC-BC-CC (and in DC-BM25, see below) can be positive without a direct citation between *i* and *j*.

DC-BM25. In this approach, direct citations are enhanced by text relations. We define the relatedness of *i* and *j*, $r_{ij}^{\text{DC-BM25}}$, as

$$r_{ij}^{\text{DC-BM25}} = \alpha r_{ij}^{\text{DC}} + r_{ij}^{\text{BM25}} \tag{5}$$

where α is a weight of direct citations relative to BM25. We obtain values of α in the following way. The average across all BM25 relatedness values greater than 0 is calculated, an average that turned out to be equal to 50. By setting α to 50, the DC values are put on the same scale as the BM25 relatedness values, in an average sense. By setting α to 25 (100), less (more) emphasis would be put on DC. We use all these three α values in our analysis.

When calculating r_{ij}^{X}, $X \in \{\text{BC, CC, BM25, DC-BC-CC, DC-BM25}\}$, we only consider the *k*-nearest neighbors to *i*, i.e. the *k* publications with the highest relatedness values with *i*. If *j* is not among the *k* publications with the highest relatedness values with *i*, $r_{ij}^{X} = 0$. Here, *k* is set to 20. For a sensitivity analysis, we refer the reader to Waltman et al. (2019). We apply the *k*-nearest neighbors technique for efficiency reasons. However, we do not apply this technique in DC, since computer memory requirements are relatively modest for DC.

Normalization of the relatedness measures and clustering of publications

For all six approaches, the corresponding relatedness measures are normalized. The *normalized relatedness* of publication *i* with publication *j* is the relatedness of *i* with *j*, divided by the total relatedness of *i* with all other publications that are considered. Now, without normalization, clustering solutions obtained using different relatedness measures but associated with the same value of the resolution parameter of the clustering (see below in this section) might be far from satisfying the requirement that compared, with regard to accuracy, solutions should have exactly the same granularity, where the *granularity* of a solution is defined as the number of publications divided by the sum of the squared cluster sizes (Waltman et al., 2017, 2019). With the indicated normalization, the granularity requirement can be assumed to be approximately satisfied by the solutions. However, to further deal with the granularity issue, granularity-accuracy plots (GA plots) are used in the study (Waltman et al., 2017, 2019). GA plots are described in the section on evaluation of approach performance below.

In this study, we use the Leiden algorithm (Traag et al., 2018a, 2018b) to generate a series of clustering solutions for each of the relatedness measures. The Leiden algorithm is used to maximize the Constant Potts Model as quality function (Traag et al., 2011; Waltman & Van Eck, 2012). Using different values of the resolution parameter γ (0.000001, 0.000002, 0.000005, 0.00001, 0.00002, 0.00005, 0.0001, 0.0002, 0.0005, 0.001, 0.002, 0.005, 0.01), we obtain 13 clustering solutions for each relatedness measure.

The normalization of the relatedness measures transforms these measures to non-symmetrical counterparts. The clustering methodology we use requires, however, that the relatedness values are symmetrical. We solve this issue in the following way. Let $\hat{r}_{ij}^{X}$ denote the relatedness of *i* with *j* with respect to approach $X \in \{\text{DC, BC, CC, BM25, DC-BC-CC, DC-BM25}\}$ after normalization of r_{ij}^{X}. The relatedness value for *i* and *j* given as input to the clustering algorithm is $\hat{r}_{ij}^{X} + \hat{r}_{ji}^{X}$, i.e. the sum of the two normalized relatedness values. Clearly, then, the relatedness values are made symmetrical before given as input to the clustering algorithm.

Evaluation of approach performance

For the evaluation of the performance of the six approaches, an external subject classification scheme, MeSH, is used. MeSH descriptors and subheadings are used to index publications in PubMed. MeSH contains more than 28 thousand descriptors that are arranged hierarchically by subject categories with more specific descriptors arranged beneath broader descriptors (National Library of Medicine, 2019). MeSH descriptors can be designated as major indicating that they correspond to the major topics of the publication, whereas non-major descriptors are added to reflect additional topics substantively discussed within the publication. Further, approximately 80 subheadings (or qualifiers) can be used by the indexer to qualify a descriptor.

We calculate a weight (information content, IC) for each descriptor (Zhu et al., 2009). Let $freq(desc_i)$ denote the frequency of descriptor *i* (here calculated over all MEDLINE publications published within the period 2013-2017), then:

$$IC(desc_i) = -\log(P(desc_i)) \tag{6}$$

where

$$P(desc_i) = \frac{freq(desc_i) + \sum_{d \in descendants(desc_i)} freq(d)}{\sum_{k=1}^{s} \left(freq(desc_k) + \sum_{d \in descendants(desc_k)} freq(d) \right)} \tag{7}$$

We then represent each publication by a vector of length $s + (s \times m)$ where *s* and *m* are the total number of unique MeSH descriptors and unique[1] number of subheadings in the dataset, respectively. The vector position for the *i*th descriptor is given by $(m + 1) \times i - m$ and the corresponding *weight for publication l* ($\omega_i(l)$) is defined as

[1] A group of MeSH descriptors that routinely are added to most articles, so called "check tags", are concepts of potential interest, regardless of the general subject content of the article (examples are "Human" and "Adult"). We do not include such check tags in any calculations.

$$\omega_i(l) = \begin{cases} 0 \text{ if } desc_i \text{ is absent in l} \\ IC(desc_i) \times 1 \text{ if } desc_i \text{ is a minor descriptor in l} \\ IC(desc_i) \times 2 \text{ if } desc_i \text{ is a major descriptor in l} \end{cases} \tag{8}$$

The vector position for the *j*th subheading connected to the *i*th descriptor is given by $(m + 1) \times i - m + j$ and the corresponding *weight for publication l* ($\phi_{ji}(l)$) is defined as

$$\phi_{ji}(l) \begin{cases} 1 \text{ if subheading } j \text{ and descriptor } i \text{ are present in } l \\ 0 \text{ otherwise} \end{cases} \tag{9}$$

Note that many descriptor-subheading pairs are nonsensical and will never exist in practice and the subheading in such a pair will thus always take on the value 0 in the vectors.

We estimate the subject similarity between the publications by the cosine similarity (Salton & McGill, 1983) between their corresponding vectors defined above. As in case of calculating relatedness in BC, CC, BM25, DC-BC-CC and DC-BM25, and for the same reason, we apply the *k*-nearest neighbors technique. As in these three approaches, *k* is set to 20. We then normalize the cosine similarities in the same way as we normalize the relatedness measures of all six approaches. Finally, the publications are clustered based on the normalized cosine similarities using the same clustering methodology, and the same set of values of the resolution parameter, as for the six approaches.

The accuracy of the *k*th $(1 \le k \le 13)$ clustering solution for $X \in \{\text{DC, BC, CC, BM25, DC-BC-CC, DC-BM25, MeSH}\}$, where the accuracy is based on MeSH cosine similarity, symbolically $A^{X_k|\text{MeSH}}$, is defined as follows (Waltman et al., 2017, 2019):

$$A^{X_k|\text{MeSH}} = \frac{1}{N} \sum_{i,j} I(c_i^{X_k} = c_j^{X_k}) \hat{r}_{ij}^{\text{MeSH}} \tag{10}$$

where $c_i^{X_k}$ ($c_j^{X_k}$) is a positive integer denoting the cluster to which publication *i* (*j*) belongs with respect to the *k*th clustering solution for *X*, $I(c_i^{X_k} = c_j^{X_k})$ is 1 if its condition is true, otherwise 0, and $\hat{r}_{ij}^{\text{MeSH}}$ the normalized MeSH cosine similarity of *i* with *j*. Recall that DC-BC-CC (DC-BM25) has two (three) variants, $\alpha \in \{1,5\}$ ($\alpha \in \{25,50,100\}$), and that we thereby, in total, work with 10 relatedness measures. Observe that we include 13 clustering solutions based on MeSH cosine similarities in the evaluation exercise. The accuracy results obtained for MeSH give an upper bound for the results that can be obtained when the relatedness measures of the six approaches are used to cluster the publications and accuracy is based on MeSH cosine similarity. We remind the reader that the value of the resolution parameter γ is held constant across the six approaches and MeSH regarding the *k*th clustering solution.

We visualize the evaluation results by using GA plots. The use of such plots is a way to counteract the difficulty that the requirement that compared, with regard to accuracy, clustering solutions should have exactly the same granularity is only approximately satisfied. In a GA plot, the horizontal axis represents granularity (as defined above), whereas the vertical axis represents accuracy. For a given approach, like DC, a point in the plot represents the accuracy

and granularity of a clustering solution, obtained using a certain resolution value of γ. Further, a line is connecting the points of the approach, where accuracy values for granularity values between points are estimated by interpolation. Based on the interpolations, the performance of the approaches can be compared at a given granularity level.

Results

In this section, we present three figures containing GA plots. The first plot contains curves for DC and the other citation-based approaches, the second for DC and the text-based approaches, whereas the last plot contains curves for all six approaches that we investigate in our analysis. As should be clear from the section "Data and methods", MeSH is consistently used as evaluation criterion. Note that all three plots contain a curve also for MeSH, where such a curve represents an upper bound for the performance of the six approaches. One might ask what the meaning, in terms of number of clusters, of different granularity levels is. When the granularity is around 0.0001, a clustering solution typically has 500 significant clusters (defined as the number of clusters with 10 or more publications). When the granularity is around 0.001 (0.01), a clustering solution typically has 5,000 (50,000) significant clusters.

The GA plot of Figure 1 visualizes the accuracy results of enhancing DC by the combination of BC and CC (α = 1, 5), as well as the performance of DC, BC and CC. CC exhibits the worst performance among the citation-based approaches. DC is outperformed by BC and the two DC-BC-CC variants, whereas BC performs slightly worse than the DC-BC-CC variants, which performs equally well.

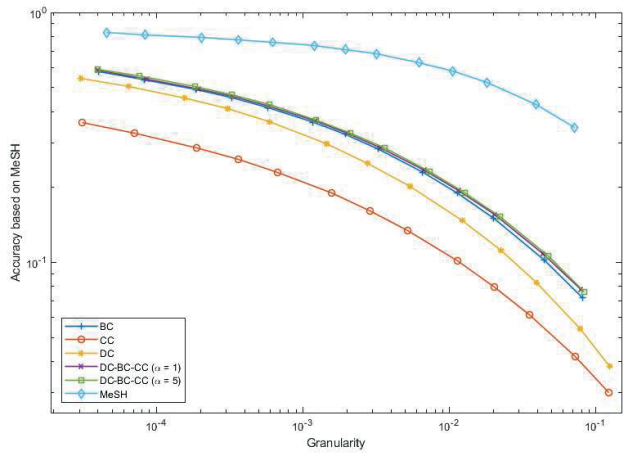


Figure 1. GA plot for comparing the approaches DC, BC, CC and the two variants of DC-BC-CC. MeSH used as the evaluation criterion.

In Figure 2, a GA plot that shows the results of enhancing DC by BM25 (α = 25, 50, 100), as well as the performance of DC and BM25, is given. BM25 performs better than DC but is outperformed by all three DC-BM25 variants. Of these, the ones with α equal to 50 and 100 perform about equally well, and better than the variant that put less emphasis on DC (α = 25).

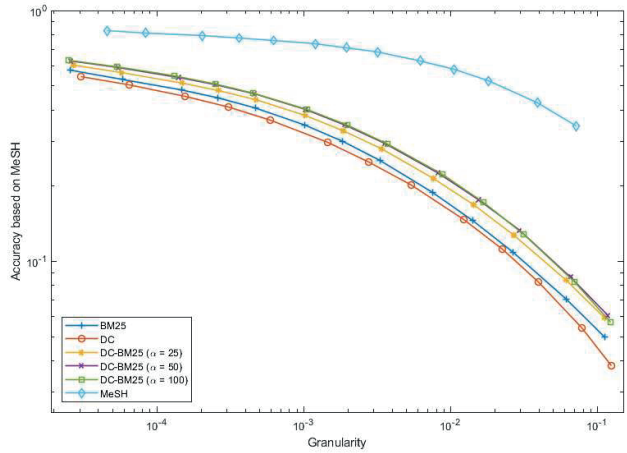


Figure 2. GA plot for comparing the approaches DC, BM25 and the three variants of DC-BM25. MeSH used as the evaluation criterion.

The GA plot of Figure 3, finally, shows the performance of DC, BC, CC, BM25, DC-BC-CC (α = 1) and DC-BM25 (α = 50). Enhancing DC by BM25 yields the best performance in our analysis, while DC-BC-CC, where DC is enhanced by the combination of BC and CC, has the second best performance, followed by BC.

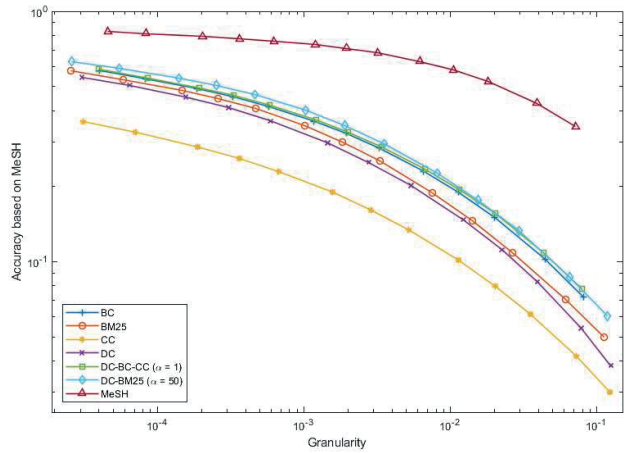


Figure 3. GA plot for comparing all six approaches. MeSH used as the evaluation criterion.

Discussion and conclusions

In this contribution, we have analyzed the effects of enhancing direct citations, with respect to publication-publication relatedness measurement, by indirect citation relations and text relations on clustering accuracy. As independent evaluation criterion, we used an approach based on MeSH, arguable the most sophisticated item-level classification scheme available. Six approaches were investigated, and the results show that enhancing DC with indirect citation relations (BC-CC) or text relations (BM25) gives rise to substantial performance gains relative to DC. The best performance was obtained by DC-BM25, followed by DC-BC-CC. Thus, in

our analysis and interestingly, enhancing direct citations by text relations gives rise to a better performance compared to enhancing direct citations by indirect citation relations.

The poor performance of CC has been observed in earlier research (Klavans & Boyack, 2017; Waltman et al., 2017, 2019) and was expected. Clearly, a publication that has not received any citations is not co-cited with another publication, and can therefore not be adequately clustered. In the study by Klavans and Boyack (2017), however, and in contrast to our results, DC yielded more accurate clusters than BC. However, we, like Waltman et al. (2017, 2019), used a more principled evaluation methodology, as well as a different evaluation criterion.

Waltman et al. (2017, 2019) compared DC, BC, CC and DC-BC-CC (α = 1, 5), using BM25 as the evaluation criterion and a considerably smaller publication set than the publication set of our analysis. Our results for these citation-based approaches (Figure 1) demonstrate the same pattern as the results of these authors. This supports the robustness of the results for the four citation-based approaches, since the two studies used different publication sets and different evaluation criteria.

In our study, BM25 performs better than DC. Boyack and Klavans (2018), though, concluded that clusters that were obtained on the basis of the text-only relatedness measures used in their study are as accurate as those that were obtained on the basis of direct citation. However, a different evaluation criterion, compared to ours, was used in the study.

Chen et al. (2017) used the TF-IDF term weighting approach combined with the cosine similarity measure in order to weight direct citations by textual similarities. We tested the same approach (without taken term position information into account), as well as an approach in which the combination of TF-IDF and the cosine measure was replaced BM25 for weighting of direct citations. These two approaches, say DC-TF-IDF and DC-BM25 (weighted links), were outperformed, though, by DC-BM25, DC-BC-CC and BC. Note that, for DC-TF-IDF and DC-BM25 (weighted links), and in contrast to DC-BM25, a necessary (but not sufficient) condition for obtaining a positive relatedness value for two publications i and j is that there is a direct citation from i to j or conversely.

A limitation of our study is that it could be argued that the MeSH approach is not fully independent of relatedness measures based on text in abstracts and titles of publications, since the indexers who assign MeSH terms to publications partially rely on the title and full-text of publications. Therefore, the MeSH approach might not be fully independent of the approaches BM25 and DC-BM25. However, MeSH constitutes a controlled vocabulary, whereas BM25 makes use of an uncontrolled vocabulary, the source of which is the authors of the publications. In view of this, we believe that the MeSH approach is sufficiently different from approaches that make use of terms in abstracts and titles.

For future research, it would be interesting to enhance direct citations by a variant of the BM25 approach, BM25F, which takes term position information into account (Craswell et al., 2005).

Acknowledgments

This work is partly funded by the National Key Research and Development Program of China (Grant No. 2017YFB1402400).

References

Boyack, K.W., Newman, D., Duhon, R.J., Klavans, R., Patek, M., Biberstine, J.R., ... & Börner, K. (2011). Clustering more than two million biomedical publications: Comparing the accuracies of nine text-based similarity approaches. *PLOS ONE*, 6(3), e18029.

Boyack, K.W. & Klavans, R. (2014). Including cited non-source items in a large-scale map of science: What difference does it make? *Journal of Informetrics*, 8(3), 569-580.

Boyack, K.W. & Klavans, R. (2018). Accurately identifying topics using text: Mapping PubMed. *Proceedings of the 23rd International Conference on Science and Technology Indicators* (pp. 107-115).

Chen, P. & Redner, S. (2010). Community structure of the physical review citation network. *Journal of Informetrics*, 4(3), 278-290.

Chen, W., Fengxia, Y. & Wang, Y. (2013). Community Discovery Algorithm of Citation Semantic Link Network. *Computational Intelligence and Design*, 2, 289-292.

Chen Y.W., Xiao X., Deng Y. & Zhang, Z. (2017). A weighted method for citation network community detection. *Proceedings of the 16th International Conference on Scientometrics and Informetrics* (pp. 58-67).

Cohn, D. & Hofmann, T. (2001). The missing link - A probabilistic model of document content and hypertext connectivity. In T. K. Leen et al. (Eds.), *Advances in Neural Information Processing Systems 13* (pp. 430-436). Cambridge: MIT Press.

Craswell, N., Zaragoza, H. & Robertson, S (2005). Microsoft cambridge at trec-14: Enterprise track. *Proceedings of the Fourteenth Text REtrieval Conference*.

Fujita, K., Kajikawa, Y., Mori, J. & Sakata, I. (2014). Detecting research fronts using different types of weighted citation networks. *Journal of Engineering and Technology Management*, 32, 129-146.

Girvan, M. & Newman, M.E.J. (2002). Community structure in social and biological networks. *PNAS*, 99(12), 7821-7826.

Hamedani, M.R., Kim, S.W. & Kin, D.J. (2016). SimCC: A novel method to consider both content and citations for computing similarity of scientific papers. *Information Sciences*, 334-335, 273-292.

Kajikawa, Y., Yoshikawa, J., Takeda, Y. & Matsushima, K. (2008). Tracking emerging technologies in energy research: Toward a roadmap for sustainable energy. *Technological Forecasting and Social Change*, 75(6), 771-782.

Klavans, R. & Boyack, K.W. (2017). Which type of citation analysis generates the most accurate taxonomy of scientific and technical knowledge? *Journal of the Association for Information Science and Technology*, 68(4), 984-998.

Kusumastuti, S., Derks, M.G., Tellier, S., Di Nucci, E., Lund, R., Mortensen, E.L. & Westendorp, R.G. (2016). Successful ageing: A study of the literature using citation network analysis. *Maturitas*, 93, 4-12.

Persson O. (2010). Identifying research themes with weighted direct citation links. *Journal of Informetrics*, 4(3), 415-422.

Ruiz-Castillo, J. & Waltman, L. (2015). Field-normalized citation impact indicators using algorithmically constructed classification systems of science. *Journal of Informetrics*, 9(1), 102-117.

Salton, G. & McGill, M. J. (1983). *Introduction to Modern Information Retrieval*. New York: McGraw-Hill.

Sjögårde, P. & Ahlgren, P. (2018). Granularity of algorithmically constructed publication-level classifications of research publications: Identification of topics. *Journal of Informetrics*, 12(1), 133-152.

Small, H. (1997). Update on science mapping: Creating large document spaces *Scientometrics*, 38(2), 275-293.

Sparck Jones, K., Walker, S. & Robertson, S.E. (2000a). A probabilistic model of information retrieval: Development and comparative experiments: Part 1. *Information Processing and Management*, 36(6), 779-808.

Sparck Jones, K., Walker, S. & Robertson, S.E. (2000b). A probabilistic model of information retrieval: Development and comparative experiments: Part 2. *Information Processing and Management*, 36(6), 809-840.

Subelj, L., Van Eck, N.J. & Waltman, L. (2016). Clustering scientific publications based on citation relations: A systematic comparison of different methods. *PLOS ON*E, 11(4), e0154404.

Traag, V.A., Van Dooren, P. & Nesterov, Y. (2011). Narrow scope for resolution-limit-free community detection. *Physical Review E*, 84(1), 016114.

Traag, V.A., Waltman, L. & van Eck, N. J. (2018a). From Louvain to Leiden: guaranteeing well-connected communities. arXiv:1810.08473 [cs.SI].

Traag, V.A., Waltman, L. & Van Eck, N.J. (2018b). CWTSLeiden/networkanalysis [Source code]. Zenodo. doi:10.5281/zenodo.1466831.

U.S. National Library of Medicine. (2019). *Introduction to MeSH*. Retrieved April 7, 2019 from: https://www.nlm.nih.gov/mesh/introduction.html.

Waltman, L. & Van Eck, N. J. (2012). A new methodology for constructing a publication-level classification system of science. *Journal of the American Society for Information Science and Technology*, 63(12), 2378-2392.

Waltman, L., Boyack, K.W., Colavizza, G. & Van Eck, N. J. (2017). A principled methodology for comparing relatedness measures for clustering publications. In *Proceedings of the 16th International Conference on Scientometrics and Informetrics* (pp. 691-702).

Waltman, L., Boyack, K.W., Colavizza, G. & Van Eck, N.J. (2019). A principled methodology for comparing relatedness measures for clustering publications. arXiv:1901.06815.

Yudhoatmojo, S.B. & Samuar, M.A. (2017). Community detection on citation network of DBLP data sample set using LinkRank Algorithm. *Procedia Computer Science*, 124, 29-37.

Zhu, S., Zeng, J. & Mamitsuka, H. (2009). Enhancing MEDLINE document clustering by incorporating MeSH semantic similarity. *Bioinformatics*, 25(15), 1944-1951.

Author Index

Printed in Italy, August 2019