An 80/20 Data Quality Law for Professional Scientometrics?

Andreas Strotmann[1] and Dangzhi Zhao[2]

[1] *andreas.strotmann@gmail.com*
ScienceXplore, D-01814 Bad Schandau (Germany)

[2] *dzhao@ualberta.ca*
University of Alberta, School of Library and Information Studies, Edmonton, Alberta (Canada)

Scientometric network error consequences

Only very recently have researchers begun looking at what concrete effect the errors in a network model caused by name ambiguities in the data sources may have on the results of popular types of network analysis. The results that they report are quite alarming in the aggregate: not only do typical evaluative analyses of individuals (e.g., citation rankings) suffer significantly from these errors, but there is mounting evidence that even the most basic statistical features of realistic large-scale networks are hugely distorted by ambiguities. Strotmann et al. (2009), for example, document significant distortions in co-authorship network visualizations, and Diesner and Carley (2013) report that "minor changes in accuracy rates of [name disambiguation] lead to comparatively huge changes in network metrics, while the set [of] top-scoring key entities is highly robust. Co-occurrence based link formation entails a small chance of false negatives, but the rate of false positives is alarmingly high."

In fact, Fegley and Torvik (2013) go so far as to dismiss one of the most famous recent results in large-scale social network analysis, the exact power-law distribution from preferential attachment (Barabási & Albert, 1999), at least in the case of scientific collaboration networks (Barabási et al., 2002), as a mere artefact produced by a lack of name disambiguation in the underlying dataset! The ultimate irony here is that Fegley and Torvik's (2013) data are consistent with an interpretation that Barabási's cooperation network power may have been induced by a power law distribution of name ambiguities rather than co-authorships.

Similarly, Strotmann and Zhao (2013) find that even highly stable statistical analysis methods of author co-citation analysis fail in the face of large-scale ambiguity errors in the underlying dataset.

While for evaluative bibliometrics the most serious problem is generally the "splitting" of individuals, i.e., the failure to recognize each and every one of an individual's contributions correctly (especially of high-performing individuals), Fegley and Torvik (2013) find that splitting is not the main concern in relational network analysis. Instead, they and Strotmann and Zhao (2013) both find that it is the erroneous "merging" of individuals, i.e., the failure to separate the contributions of multiple individuals correctly because their names are too similar, that causes major distortions of large-scale network analysis results in relational network analysis. Especially East Asian names are prone to extreme amounts of merging. While in European cultures there are relatively few common given names but a large variety of family names, in Chinese, Korean and other East Asian cultures the opposite is the case—a small number of surnames is shared by half their populations, but given names are much more varied. The old tradition in scientific publishing to list authors by their surnames and initials works, sort-of, when science is done in European-origin cultures, but all bibliographic databases have in recent years had to move to a full-name model as research boomed in the Asian Tiger nations (e.g., PubMed/MEDLINE in 2002).

When is a scientometric network sufficiently complete and clean?

As Torvik and Smalheiser (2009) make abundantly clear, it is for all intents and purposes impossible to disambiguate the names of all the individuals in a large dataset completely and fully correctly. With absolute perfection thus out of the question, what remains is to ask when a disambiguation is "good enough", and if (and how) it is possible for a typical researcher to go about disambiguating the dataset well enough. Unfortunately, there is very little research, if indeed any, into what constitutes "good enough" for a scientometric study. The few studies that have looked into what goes wrong when individuals are not recognized correctly do give us a hint, though.

First of all, "good enough" usually means that the most important contributions of the top-ranked individuals must be absolutely correctly attributed. Whatever other good methods (e.g., name disambiguation algorithms or author registries) we may find to disambiguate our data, in the end it will therefore be necessary to manually double-check, and where necessary fix, the highest-impact individuals' data. Secondly, some statistical procedures or network measures are more vulnerable than others to name ambiguities. Local network measures (e.g., node degree) are less affected than global ones (e.g., size of connected component), and evaluative studies (e.g., ranking) are more affected than relational ones (e.g., correlations) (Diesner & Carley, 2013; Strotmann & Zhao, 2012).

An 80/20 scientometric data quality rule?

For ranking studies, absolute correctness is paramount, and huge efforts need to be expended to get all the top-ranked individuals just right. When the "individuals" are research institutions, this can be a daunting task. For correlative studies, on the other hand, a study by Albert, Jeong, and Barabási (2000) warns us that, while global measures of power-law distributed networks may be quite resilient to *uniformly* distributed random errors, they are also quite vulnerable to the kind of *highly skewed* error distributions that we observe for name ambiguities, for example. In the case of an extremely skewed error distribution, they observed that an error rate as low as 10%-20% completely changed the measured values for a fundamental global network metric, namely, connectivity.

We can take this as a warning that, as a rule of thumb, we generally need to aim for a roughly 90% (but definitely 80% or better) complete and correct dataset when error distributions are skewed. Note that the requirement of 80% completeness or better applies, in particular, to the underlying citation index's coverage of the field being studied: a focus on high-impact literature implies a highly skewed error distribution! On the plus side, studies on the life sciences can thus be relied upon to yield reliable results as long as their disambiguations are good. Results from *any* scientometric study on the social sciences, however, are suspect as long as they rely on these databases and these databases cover much less than 80% of the literature in those fields.

Note that an 80% data correctness requirement for a professional scientometric study would apply to the data as it is used for network statistics. When both data collection *and* cleaning are subject to highly skewed error distributions, this means that we need 90% correct data collection *and* 90% correct data cleaning to guarantee 80% correct data for analysis.

Conclusions: the bad news and the good

This, then, is the bad news for those who aim to provide a truly professional scientometric service to their community: power-law-like data *and* error distributions may mean that only nearly-complete *and* nearly-clean datasets can be trusted to serve as a reliable basis for nearly *any* type of network or statistical analysis.

The good news is that there are plenty of successful bibliometric studies that imply that this level of correctness is also usually quite sufficient for meaningful studies, as long as only "local" measures or relational statistics are required. There *are* fields that are covered to 90%+ in citation databases, e.g., the citable literature of the life sciences, and there *are* disambiguation methods (e.g., some of those reviewed in Smalheiser & Torvik, 2009 or that of Strotmann et al., 2009) that do make reliable scientometric studies possible. However, scientometric professionalism may well require that these methods be utilized in nearly *all* future studies, and thus, that they be applied to, and adopted by, the citation databases themselves.

Acknowledgments

This Ignite Talk extends, with permission of the publisher, Section 4.4, "Disambiguation in Citation Network Analysis: Ambiguity and Power Laws," of Zhao & Strotmann (2015).

References

Albert, R., Jeong, H.W. & Barabási, A.L. (2000). Error and attack tolerance of complex networks. *Nature 406*, p.378

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

Barabási, A.L., Jeong, H., Neda, Z, Ravasz, E, Schubert, A. & Vicsek, T. (2002). Evolution of the social network of scientific collaborations. *Physica A, 311*, p. 590

Diesner, J. & Carley, K.M. (2013). Error propagation and robustness of relation extraction methods. *XXXIII International Sunbelt Social Network Conference*, Hamburg, Germany, May 2013.

Fegley, B.D. & Torvik, V.I. (2013). Has large-scale named-entity network analysis been resting on a flawed assumption? *PLoS ONE 8* (7): e70299.

Smalheiser, N. R. & Torvik, V. I. (2009). Author name disambiguation. *Annual Review of Information Science and Technology* 43, 287.

Strotmann, A., Zhao, D. & Bubela, T. (2009). Author name disambiguation for collaboration network analysis and visualization. *Proceedings of the American Society for Information Science and Technology 2009 Annual Meeting*, November 6–11, 2009, Vancouver, BC, Canada

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), p.1820

Torvik, V. I. & Smalheiser, N. R. (2009). Author name disambiguation in MEDLINE. *ACM Transactions on Knowledge Discovery from Data, 3* (3)

Zhao, D. & Strotmann, A. (2015). *Analysis and Visualization of Citation Networks.* Morgan & Claypool.