Patterns of International and National Web Inlinks to US University Departments: A Webometric Analysis of Disciplinary Specificity

Rong Tang[1]
School of Information Science and Policy, State University of New York at Albany, 113 Draper Hall, 135 Western Avenue, Albany, NY 12222
E-mail: *tangr@albany.edu*

Mike Thelwall
School of Computing and Information Technology, University of Wolverhampton, Wulfruna Street, Wolverhampton WV1 1SB, UK. E-mail: m.thelwall@wlv.ac.uk

Abstract

An investigation of links to 89 US academic departments from three different disciplines gave insights into the kinds of international regions and national domains that linked to them. Here we describe our data collection and analysis procedure. While we found significant correlations between total counts of international inlinks and total publication impact in Psychology and Chemistry, counts of international inlinks to History departments were too small to form a significant correlation. The correlations suggest that international links may provide evidence of scholarly communication patterns. History departments attracted a significantly lower percentage of international inlinks than to those of Chemistry and Psychology, but the main source of links for all three disciplines was from Europe. Analyses of mainly national inlinks, characterized by gTLDs (generic Top Level Domains), showed that the major source of links for all disciplines was .edu sites, followed by .com, .org, .net. As a whole, regional disciplinary differences were stronger than gTLD differences, although both differences were small. This could be surprising, given the choice of a hard science, a social science and a humanities discipline.

1 Background

The Internet has forever changed the landscape of higher education. Web technology bestows higher education institutes never-before capacities for "the creation of learning communities that defy the constraints of time and distance as it provides access to knowledge that was once difficult to obtain…the Internet enables education to occur in places where there is none, extends resources where there are few, expands the learning day, and opens the learning place" (Web-Based Education Commission, 2000, p.i-iii.). With the emergence of distance learning, educational portals and virtual universities, the nature of higher learning has been redefined, the format of course instruction, knowledge delivery and dissemination has been transformed. All of this has fostered a world-wide education network and a global knowledge society. As a result, Web based studies are starting to address similar questions to those that were previously

[1] To whom all correspondence should be addressed

Rong Tang and Mike Thelwall, *Patterns of international and national web inlinks to US university departments: a webometric analysis of disciplinary specificity*, In: Guohua Jiang, Ronald Rousseau, Yishan Wu (Eds), Proceedings of the 9th International Conference on Scientometrics and Informetrics – ISSI 2003, Dalian University of Technology, Dalian, China, 2003, 312-320.

exclusively in the domain of citation analysis (e.g. Lange, 1985; Glänzel & Schubert, 2001; Borgman & Furner, 2002).

The Web, especially hyperlinks in university Web sites, has been heralded as an important data source for information scientists (Almind & Ingwersen, 1997; Rousseau, 1997; Aguillo, 1998; Ingwersen, 1998; Cronin, 2001; Borgman & Furner, 2002). Links have been mined by researchers mainly for the purpose of discovering patterns of online communication between universities. Previous Webometric studies have analysed groups of universities (Smith, 1999; Smith & Thelwall, 2002; Thelwall, 2002a), departments (Thomas & Willett, 2000; Thelwall & Tang, 2002; Li, Thelwall, Musgrove, & Wilkinson, 2003) or journal web sites (Smith, 1999; Vaughan & Thelwall, 2003), but only the predecessor to this paper has focused upon disciplinary differences within a set of departments from different subjects (Tang & Thelwall, 2003). Many studies have found that link counts positively correlate with the universities' or departments' academic status and research performance, although such links do not tend to be created for reasons analogous to journal citations (Wilkinson, Harries, Thelwall, & Price, 2003). International linking patterns have also been described and analysed in some recent articles, both for the Asia-Pacific region (Thelwall & Smith, 2002) and Europe (Polanco, Boudourides, Besagni, & Roche, 2001; Thelwall, Tang, & Price, 2003).

This study is designed to investigate the patterns of inlinks to a total of 89 US departments in the fields of Chemistry, Psychology, and History. This is a continuation of our recently completed study that focuses on the interlinking pattern *between* departments (Tang & Thelwall, 2003). In this study, examinations are made both on regional differences among international inlinks and URL domains (.com, .edu, .net, etc.), and differences between the proportion of national and international inlinks to more than 23 departments from each discipline. An overall disciplinary comparison is made based on analysis of the number and kinds of international and national Web sites that the departments of a discipline appear to attract.

2 Research Questions

The first research question concerns validating the link count data as potentially usefully information by comparison with another existing measurement, i.e. ISI publication counts coupled with citation impact factors. Positive results would not provide conclusive evidence of value, but would be suggestive of the merit of further link analysis. The remaining questions are related to patterns of international and national inlinks, as specified below:

1. Do the counts of international inlinks correlate with ISI publication impact values?
2. Which regions of the world most commonly link to the selected set of US academic departments of three disciplines? Secondly, are there regional differences among international inlinks? Specifically, do departments from a particular discipline tend to be frequently targeted by particular international regions?
3. Are there disciplinary differences with regard to the percentage of international and national inlinks?
4. Are there URL-domain differences among national inlinks? Specifically, (a) do departments from a particular discipline tend to be frequently targeted by Web

sites from particular generic Top Level Domains (gTLDs), (b) do links from particular domains correlate with research ratings for each set of departments and gTLD, and (c) is the pattern of inlinks from one gTLD to a set of departments within a discipline similar to the pattern of inlinks from another gTLD, for each of the three disciplines and all pairs of gTLDs?

3 Methodology - Data Collection

Data collection was performed through the AltaVista "Advanced Search" Option. AltaVista was chosen rather than Google, which has greater web coverage, because of the greater range of search options available for its advanced searches.

3.1 International Inlinks All inlink counts are generated by typing the `Link:` command in the "Search with … this Boolean Expression" box. To get the specific inlink counts from a particular region, select the relevant region from the "location: by regions" menu. For instance, to search Asian inlinks to University of Arizona Chemistry Department. The following command is issued:

```
LINK:chem.arizona.edu
```

And select "Asia" from the location menu. For the North American region we modified the command to exclude links from the same university. In each case the extra requirement AND NOT HOST:.. was used with the domain name of the whole university. For example in the above case the command was:

```
LINK:chem.arizona.edu AND NOT HOST:arizona.edu
```

Note that unfortunately after the study was conducted, AltaVista produced a new search interface and its regional search facility was no longer available.

3.2 National Inlinks National inlink counts from different URL domains are generated in two-steps: 1. Type the second Boolean expression above and select "North America" from the region menu. 2. With the same command and type ".ca" (Canada) in the "Location: by Domain" box. 2. With the same command and type ".mx" (Mexico) in the "Location: by Domain" box. The US counts should be subtracting counts from No.2 from No.1. In other words, all North American site inlinks except those from outside the US (i.e. Mexico and Canada). Note that in most cases the counts from No. 2 and No. 3 were 0.

3.3 gTLD Inlinks Counts of generic Top Level Domain inlinks were obtained with the same procedure as described above, except that the query was modified by adding "AND DOMAIN:edu" for a restriction to links from edu sites, and an equivalent modification for links from com, org, net, mil, int and gov sites. These are the original TLDs. There are now newer ones in operation, including .biz and .info, and these were not used because their newness would mean that they would be unlikely to yield useful results.

3.4 Research ranking and Faculty Size Research ranking of the 89 departments was established in our previous study. The research index that we used is called publication impact value. Below is a brief recount of the data collection procedure:

Top 30 Departments Ranked based on Year 2000 Publications. For the three disciplines, we searched Chemistry via ISI Social Science Citation Index (dialog file 34),

Psychology via Social Science Citation Index (dialog file 7), and History via ISI Arts and Humanities Citation Index (dialog file 439). We only included year 2000 publication counts because at the time of the data collection, the year 2002 data was incomplete, and the year 2001 was not considered because the concern about September 11 event influencing the normal publication pattern. Searches were conducted via dialoglink and the command procedure is as follows, with separate search commands for each discipline included in the brackets:

S SC=chemistry [SC=psychology, SC=history]
S CS=(dept? (3N) chem?) [CS=(dept? (3N) psychol?), CS=(dept? (3N) hist?)]
S PY=2000 and GL=USA
S S1 AND S2 AND S3

This allows the search to be limited to a particular subject category with publications from a particular department and publication year as 2000 and location as US. After this, a rank command is issued: RANK CS. In issuing the rank command, publication counts in Chemistry was over 10,000, which exceed the limit of ranking set. Two rank commands were then issued to perform the ranking for the first 10,000 and then another for the remaining items. The top 100 ranked departments from both lists were merged and numbers were added to produce the entire ranking for Chemistry.

The CS ranking function in dialog may produce errors in data, due to the same department being indexed with a different name and thereby being given a different entry in the ranking. To compensate for this, we selected the top 100 CS for each discipline, double-checked for duplicated entries and made corrections for those that were in fact the same department/institute but were counted separately. In the end, the total number of departments became 89 after searching the department's Web site. Initially we listed more than 30 departments for all disciplines but not all departments had Web sites, especially the ones in History, and so we ended up with 24 in History, 32 in Psychology and 33 in Chemistry.

ISI Citation Impact Factor. Since we were more interested in the sitation counterpart of the bibliometric indicators, we decided to get specific citation counts for each department in the selection. We contacted ISI Contract Research Services and obtained the citation Impact Factor (i.e. average citations per article) for the year 1997-2001 for universities (not necessarily from only the particular department that was selected) by the fields of Chemistry, Psychology, and History.

Publication Impact Value of Departments. Given that the ISI citation impact data was oriented to field in a university and not necessarily the particular department in that university, we decided to multiply the number of publications from a department with the citation impact factor of a given field from the university of that department. This creates a count that we operationally defined as the "Publication Impact Value" of a department. For instance, suppose the publication counts in the year 2000 for the Department of Psychology at University of Illinois at Urbana-Champaign is X, and the ISI citation impact factor in the field of Psychology for University of Illinois at Urbana-Champaign is Y. The PIV for Department of Psychology at University of Illinois at Urbana-Champaign is

PIV = X * Y

A new ranking was thus generated for each discipline based on the citation value of the top 25-30 departments in that discipline. We also used the faculty size data that we collected for our previous study. The data were obtained from national professional organizations such as American Historical Association, American Chemical Society, and American Psychological Association, as well as from individual departments.

4 Results and Discussion

4.1 International Inlinks - Correlations with Research Ratings

With regard to departments from *Chemistry*, we found a significant correlation between total international inlinks and total publication impact (Spearman, r=0.353, n=33, significant at the 0.05% level). No significant correlation between international inlinks per faculty member and total publication impact per faculty member (Spearman, r=0.298, n=33, significant at the 0.10% level) was found. Although the r-values in the correlations are similar, they fall on either side of the 5% level significance threshold.

With regard to departments from *Psychology*, we found significant correlation between total international inlinks and total publication impact (Spearman, r=0.443, n=32, significant at the 0.05% level). We also found a significant correlation between international inlinks per faculty member and total publication impact per faculty member (Spearman, r=0.416, n=32, significant at the 0.05% level).

Link counts for departments in *History* were too small to permit valid statistical analysis.

4.2 The Regional Spread of International Inlinks

Figure 1 below displays the regional distribution of international inlinks to the departments of the three disciplines. We are concerned with the relative size of the different regions in this case, rather than the total number of links for each set of departments.

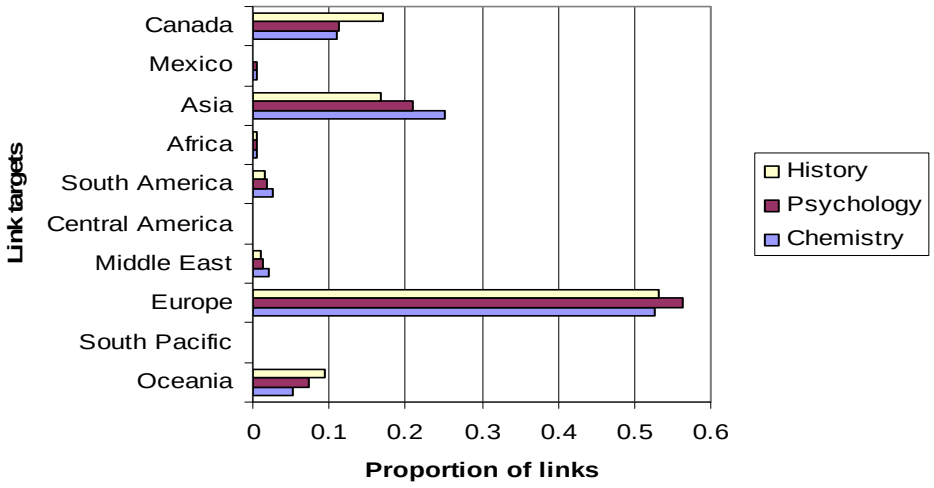

Figure 1. Inlinks to each of the three types of US department, reported by region of origin.

As can be seen from above figure, the proportions of links coming from each country or region are similar for each discipline. Europe is the main source of links although Asia, Canada and Oceania are also evident. It is somewhat surprising that apart from Canada, the rest of the continent of America contributes relatively few links. In academic cyberspace the USA seems to be closely connected with Europe and Asia.

Within a particular region, History has the highest proportion of links for Canada and Oceania. In Asia, the highest proportion is for Chemistry links whereas in Europe, Psychology has the highest proportion of links. This suggests that different international regions or countries are attracted to different disciplines in the domain of Science, Social Science, and Arts and Humanities. It appears that Western countries are relatively more attracted to Arts and Humanities as well as Social Science departments in the US, while the main attention from Eastern countries and South America is on Science departments. Readers are warned that the proportional differences may not be significant and comparisons across regions may be imprecise due to the unreliability of the data source (Rousseau, 1999; Snyder & Rosenbaum, 1999; Bar-Ilan, 2001; Björneborn & Ingwersen, 2001) and methodological problems (Thelwall, 2003).

4.3 Disciplinary Differences in the Proportion of International to National Inlinks

The proportion of inlinks that were international was: Chemistry, 19%; Psychology, 16%; and History, 6%. Clearly History attracts relatively low proportion of international links. This may reflect the fact that History as a subject domain is partially cultural-dependent since some branch is concerned with matters that are primarily of national interest: a chronicle of the host country. This is a conjecture that would be difficult to test for empirically but it would be interesting to repeat the analysis for other humanities disciplines that do not have a logical connection to a partially nation-specific remit.

5 The gTLD Spread of All Inlinks

5.1 Inlink Proportions by gTLD

Figure 2 illustrates the spread of site inlinks by top-level domain of origin. From this it can be seen that there is are no significant disciplinary differences. The graph brings no surprises: other educational institutions are a major source of links, as are the com, org and net domains. The latter three contain a variety of types of sites, not just commercial, non-commercial organisations, and network-oriented sites as their names suggest. For example, “The .org domain is ... intended to serve the non-commercial community, but all are eligible to register” (IANA, 2003). In fact they are open to anyone who wishes to buy them and so to form a meaningful conclusion about the implication of the relative size of the bars in Figure 2, we would have had to classify a random sample of the linking pages as to organization type. However, classification is known to be a difficult exercise on the Web (Furner, Ellis, & Willet, 1999; Wilkinson, Harries, Thelwall, & Price, 2003) and we have not attempted this. Nevertheless, it is clear both that US academic institutions account for at least half of all links, and US government bodies only a tiny amount.

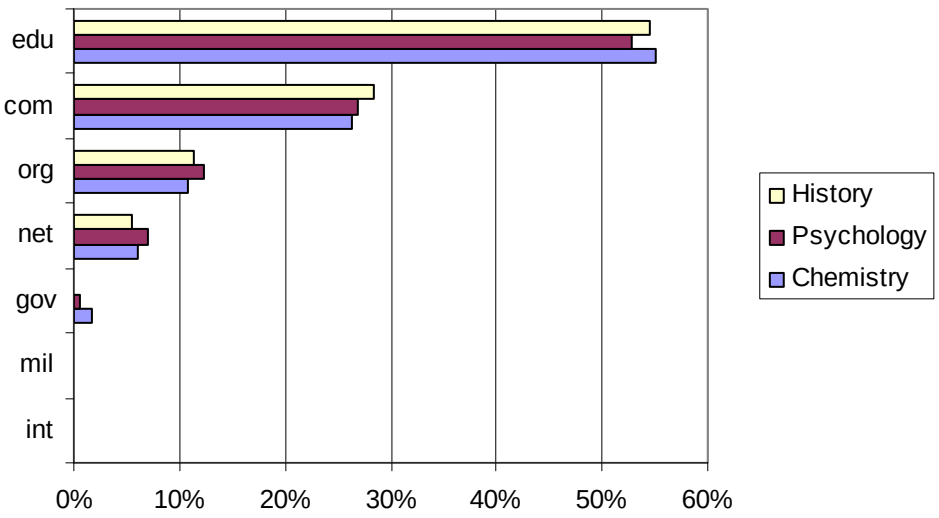

Figure 2. Inlinks to each of the three types of US department, reported by domain of origin.

5.2 Inlink Association with Research Productivity, by gTLD

We calculated Spearman correlation coefficients between inlinks and research productivity for all of the domains except int, which hosted no links to any department. For Chemistry, all gave positive results, of similar magnitudes, but only two were over the threshold for significance. A similar scenario was true for Psychology departments, except that the correlations were mostly stronger, particularly for edu inlinks. For History, there were no significant correlations with publication impact. High values in the table would indicate subject-gTLD pairs where more productive departments attracted more links from the gTLD. This could perhaps be either as a result of more productive departments engaging more with, or being more attractive to, sites in the gTLD. The differences are not really large enough to draw firm conclusions, however.

Table 1. Spearman correlations between inlinks and publication impact ratings for the departments in the three subjects (* significant at the 5% level).

Subject	GOV	NET	ORG	COM	MIL	EDU
Chemistry	0.30	0.37*	0.33	0.35*	0.20	0.33
Psychology	0.24	0.40*	0.40*	0.33	0.31	0.48*
History	0.03	0.16	0.10	0.14	0.21	0.15

5.3 Testing for Splits in gTLD inlink attraction within Disciplines

If within a discipline one set of departments specialized in pure research and another in applied research then we would perhaps expect that links from the .com domain would not correlate with links from the .edu domain, with each being high for only one of the two types of departments. To test for this kind of dichotomy, we correlated inlink counts with each other, and each pair, and for each discipline (6X5/2=15 per discipline). This gave significant results at the 1% level in every case (excluding .int

inlinks). For example .gov inlinks for Chemistry departments correlated significantly with .com inlinks. Even for History, all pairs of inlinks correlated significantly with each other at the 1% level, with values ranging from 0.53 (between .mil and .net) to 0.89 (between .com and .net). This provides evidence that the online impact of these departments is quite universal, without a significant tendency for something like a dichotomy between a subset of departments within a discipline having an impact in a different sector to the rest within that discipline.

6 Conclusions

This is the first study that investigates the distribution patterns of international inlinks to sets of US departments. The results provided significant evidence of an association between traditional research productivity and online visibility, measured by international inlinks to the US university departments. This is reassuring evidence that the study of such links may be of value to those interested in the behaviour of researchers: links are not created completely at random.

The analysis of regional sources of links to the departments showed only a little variation by discipline. The majority of links in each case came from Europe. The pattern is probably that higher Web using nations create more links. Nevertheless, the small number of links from Mexico and Central and South America is surprising, given evidence for geographic proximity as a factor in academic Web link creation (Thelwall, 2002b).

The one area in which clear disciplinary differences did emerge was in the comparison of the proportion of national to international links, with History attracting a very low percentage of international inlinks. This probably reflects at least a proportion of history research in the US being concerned with US-specific events. Having said that, the proliferation of American studies courses internationally means that US history is perhaps of more interest internationally than that of most other countries.

The main conclusion from the analysis of gTLD sources of inlinks was that the results from different sources correlated very significantly with each other, showing that for the set of departments chosen, there was no trend for groups of specializations that would be particularly attractive for one constituency, e.g. commercial or educational.

In summary, we were able to obtain some useful information from the AltaVista link data. Since there was evidence that online patterns of linking reflect research productivity, similar (and more reliable) results could probably have been obtained from a citation analysis using the Institute for Scientific Information databases (e.g. Glänzel & Schubert, 2001), but this would have taken considerably more effort for the data collection stage. In particular, it is difficult to identify the national origin of the authors of scientific papers automatically and for a large number. If such an exercise were to be undertaken then it would be extremely fruitful to compare the results in detail with those from our study to identify the ways in which patterns of formal scholarly communication were significantly different from those of the informal online communication measured here.

As a final point, the main advantage of the Webometric over the traditional scientometric approach is its accessibility and ease of use and the disadvantage is the lower reliability of the raw data used. It also measures a different, but related, aspect of

scholarly communication. It is therefore best used for exploratory studies rather than evaluative ones, and is also highly suitable for use in triangulated investigations.

References

Aguillo, I. F. (1998). STM information on the Web and the development of new Internet R&D databases and indicators. In: *Online Information 98: Proceedings*. Learned Information, 239-243.

Almind, T. C., & Ingwersen, P. (1997). Informetric analyses on the world wide Web: methodological approaches to 'Webometrics'. *Journal of Documentation*, 53(4) 404-426.

Bar-Ilan, J. (2001). Data collection methods on the web for informetric purposes - A review and analysis, *Scientometrics.* 50(1) 7-32.

Björneborn, L., & Ingwersen, P. (2001). Perspectives of webometrics, *Scientometrics*, 50(1), 65-82.

Borgman, C., & Furner, J. (2002). Scholarly communication and bibliometrics. In: Cronin, B. (ed.), *Annual Review of Information Science and Technology 36,* Medford, NJ: Information Today Inc, pp. 3-72.

Cronin, B. (2001). Bibliometrics and Beyond: Some thoughts on web-based citation analysis. *Journal of Information Science*, 27(1), 1-7.

Furner, J., Ellis, D., & Willett, P. (1999). Inter-linker consistency in the manual construction of hypertext documents, ACM Computing Surveys (CSUR), 31(4es) [On-line supplement].

Glänzel, W., & Schubert, A. (2001). Double effort = Double impact? A critical view at international co-authorship in chemistry, Scientometrics, 50(2), 199-214.

IANA (2003). Generic Top-Level Domains. Retrieved 8 January, 2003 from http://www.iana.org/gtld/gtld.htm

Ingwersen, P. (1998). The calculation of Web Impact Factors. *Journal of Documentation*, 54(2), 236-243.

Kling, R., & McKim, G. (2000). Not just a matter of time: field differences in the shaping of electronic media in supporting scientific communication. *Journal of the American Society for Information Science*, 51(14), 1306-1320.

Lange, L. (1985). Effects of disciplines and countries on citation habits. *Scientometrics*, *8*(3-4), 205-215.

Li, X., Thelwall, M., Musgrove, P., & Wilkinson, D. (2003, to appear). The relationship between the links/Web Impact Factors of computer science departments in UK and their RAE (Research Assessment Exercise) ranking in 2001. *Scientometrics.*

Polanco, X, Boudourides, M. A., Besagni, D., & Roche, I. (2001). Clustering and mapping Web sites for displaying internal associations and visualising networks. University of Patras. Retrieved 14 January, 2003 from http://www.math.upatras.gr/~mboudour/articles/web_clustering&mapping.pdf

Rousseau, R. (1997). Sitations: an exploratory study, *Cybermetrics*, 1. Available: http://www.cindoc.csic.es/cybermetrics/articles/v1i1p1.html.

Rousseau, R., (1999). Daily time series of common single word searches in AltaVista and NorthernLight, *Cybermetrics*, 2/3. Available: http://www.cindoc.csic.es/cybermetrics/articles/v2i1p2.html

Smith, A. G. (1999). A tale of two web spaces: comparing sites using Web Impact Factors, *Journal of Documentation*, 55(5), 577-592.

Smith, A., & Thelwall, M. (2002). Web Impact Factors for Australasian universities, *Scientometrics*, 54(3), 363-380.

Snyder, H., & Rosenbaum, H. (1999). Can search engines be used for web-link analysis? A critical review. *Journal of Documentation*, 55(4), 375-384.

Tang, R., & Thelwall, M. (2003, to appear). Disciplinary differences in US academic departmental Web site interlinking, *Library & Information Science Research*.

Thelwall, M., & Smith, A. (2002). A study of the interlinking between Asia-Pacific university Web sites, *Scientometrics* 55(3), 363-376.

Thelwall, M., Tang, R., & Price, E. (2003). Linguistic patterns of academic Web use in Western Europe, *Scientometrics*, 56(3), 417-432.

Thelwall, M., & Tang, R. (2002). Disciplinary and linguistic considerations for academic Web linking: An exploratory hyperlink mediated study with Mainland China and Taiwan, University of Wolverhampton.

Thelwall, M. (2002a). Conceptualizing documentation on the web: an evaluation of different heuristic-based models for counting links between university web sites, *Journal of the American Society for Information Science and Technology*, 53(12), 995-1005.

Thelwall, M. (2002b). Evidence for the existence of geographic trends in university web site interlinking, *Journal of Documentation*, 58(5), 563-574.

Thelwall, M. (2003, to appear). Methods for reporting on the targets of links from national systems of university web sites. *Information Processing and Management*.

Thomas, O., & Willett, P. (2000). Webometric analysis of departments of Librarianship and information science. *Journal of Information Science*, 26(6), 421-428.

Vaughan, L., & Thelwall, M. (2003). Scholarly use of the web: What are the key inducers of links to journal web sites? *Journal of the American Society for Information Science and Technology*, 54(1), 29-38.

Wilkinson, D., Harries, G., Thelwall, M., & Price, E. (2003). Motivations for academic Web site interlinking: Evidence for the web as a novel source of information on informal scholarly communication, *Journal of Information Science*, 29(1), 59-66.

A revised version has been published as:

Rong Tang and Mike Thelwall (2004). Patterns of national and international Web inlinks to US academic departments: an analysis of disciplinary variations. Scientometrics, 60(3), 475-485.