A university-centred European Union link analysis

MIKE THELWALL,[a] ALESIA ZUCCALA[b]

[a] *School of Computing and Information Technology, University of Wolverhampton, Wolverhampton (UK)*
[b] *Science System Assessment (SciSA) Unit, Rathenau Instituut, Den Haag (The Netherlands)*

University web sites play an important role in facilitating a wide range of types of communication. This paper reports an analysis of international academic linking in Europe, with particular reference to European Union (EU) integration. The Microsoft search service was used to calculate international interlinking to universities and from universities. Four different web topologies were found for the link structure data and poorly connected countries were identified. The results show the expected EU dominance of the large richer Western European nations, particularly the UK and Germany. The new EU countries are not yet integrated into the EU web but some show strong regional connections.

Introduction

Although the primary outputs of science are academic papers, with an increasing role for commercial activities such as patenting and technology transfer, university web sites perform a range of useful supporting roles. These include advertising research projects results and capabilities as well as marketing educational opportunities and providing information for current staff and students [MIDDLETON & AL., 1999].

The web may be particularly important to help build international connections. For example, it seems that maintaining an up-to-date web site is sometimes seen as an essential research group activity [FRY, 2006], perhaps because it can help in recruitment as well as in general research publicity [BARJAK, 2005]. Hence for organisations such as the European Union (EU) that promote academic integration, ensuring that academic web publishing is universally effective is a desirable goal. Counts of hyperlinks to a web site or larger web space have been proposed as a web impact measure [INGWERSEN, 1998] and previous research (reviewed below) suggests that it is reasonable to use international links to investigate the extent of academic web integration and the impact of individual nations' university web sites within a region.

A few reported hyperlink studies have focussed on international academic web interconnectivity. In a study of 16 European countries, the importance of English for academic interlinking was established, accounting for half of the international linking pages. In addition, countries' universities tended to link most to countries with shared languages or which were geographically close [THELWALL & AL., 2003]. Similar patterns may also appear within a single country such as the UK (geographic) [THELWALL, 2002] and Canada (linguistic/cultural and geographic) [VAUGHAN, 2006; VAUGHAN & THELWALL, 2005]. Greece was an exception in the EU study: publishing predominantly in Greek and being almost disconnected from the rest of the (then) EU. A parallel study of the universities in the 15 European Union Member states (before 2004) confirmed that international linking is associated with country size [HEIMERIKS & VAN DEN BESSELAAR, 2006], and that there was some clustering amongst countries, although not in a completely intuitive way. Also in Europe but focussing on life sciences research groups, an EU-funded research project identified the UK and Germany as central actors on the web, reflecting their research and strengths in the field, but with France being relatively marginal online [THELWALL & LI, 2005]. At the level of research teams within Europe, web linking does not seem to directly reflect patterns of collaboration or co-authorship, at least for biotechnology, artificial intelligence and information science, and web sites appeared to be a "social and public interface" for research [HEIMERIKS & AL., 2003]. A follow-up study found that the international interlinking of research teams was significantly more cognitively structured than the international interlinking of universities, presumably because the latter is partly an aggregation of different cognitive structures [HEIMERIKS & VAN DEN BESSELAAR, 2006]. On an intercontinental scale, linking between the top 5 universities in each ASEAN country (Asia and Europe) showed a European dominance with the UK and Germany again leading [PARK & THELWALL, 2006]. Within the Asia-Pacific area, the links between universities aggregated on the national level seemed to reflect international patterns of scientific collaboration, except that smaller countries seemed to be over-represented on the web, perhaps because of exhaustive web sites that linked to every country/university/department matching a given criterion [THELWALL & SMITH,

2002]. A study of the international links between departments of biology, chemistry and physics in Australia, Canada and the UK found significant differences at both the national and the disciplinary level [LI & AL., 2005A, 2005B].

Despite the above-reviewed research, new investigations are needed into the EU academic web. This is because previous research has not covered the recently expanded EU, which has a current need for rapid integration. Moreover, international academic link analysis is dependant upon the facilities provided by commercial search engines, and these have recently changed with the introduction of Yahoo! and Microsoft APIs (Applications Programming Interfaces), which are more powerful for link analysis than the existing Google API [MAYR & TOSQUES, 2005]. In this paper we explore the potential of current link analysis techniques for assessing EU academic web integration, focussing on two research questions.

1. *Web presence*: Are there any countries in Europe that are falling behind in the use of the web for universities?
2. *Topology*: What is the overall structure of the academic web in Europe?

Link analysis methods

Search engine queries

There are currently two sources of information about links between web sites: commercial search engines and personal crawlers. A personal crawler, also known as a robot or spider, is a computer program that downloads web pages and follows their links recursively. For example it might be fed with the URL of a web site's home page and then attempt to find and download all of the remaining site pages. Researchers have used web crawlers to investigate university web site interlinking. The scale of most such research has been national [THELWALL & AL., 2005] because crawling university web sites is time consuming. Two of the major families of search engines, Yahoo! and MSN (The Microsoft Network search engine, also known as Windows Live) can also be used for link counting using their `linkdomain:` and `site:` advanced search commands. For example `linkdomain:wlv.ac.uk site:knaw.nl` returns pages within the knaw.nl domain (including subdomains like niwi.knaw.nl) that contains links to the wlv.ac.uk domain (again including subdomains). Hence it is relatively easy to count the number of links between any pair of web sites (actually link pages rather than links, but we blur this distinction in the remainder of the paper) using MSN or Yahoo! There is a key limitation, however: search engine coverage of web sites is often incomplete and their reported total number of results can be estimates, and hence the figures can be inaccurate.

One previous study has used the commercial search engine AltaVista (now joined with Yahoo!) to count links between university web sites throughout Western Europe. This study constructed large composite Boolean queries by combining many separate

queries together. For example, one long query returned pages in any Belgian university that linked to any Greek university [THELWALL & AL., 2003]. This method is no longer practical, however, because testing for this research revealed that search engines now return incorrect results for very long queries. In theory, it is possible to count the number of link pages from all the universities in the country A to all of the universities in country B through a separate query for each pair of universities but this is impractical if too many universities are involved.

Two similar queries are more practical. It is possible to count the pages linking between a university and a country code domain and this can be achieved with a single query. For example `linkdomain:wlv.ac.uk site:de` counts pages in the German (de) domain that link to the University of Wolverhampton. Similarly, `linkdomain:de site:wlv.ac.uk` counts University of Wolverhampton pages that link to the German domain. It is possible to measure the connectivity between the 149 UK universities (or university colleges) and the German domain by submitting 298 queries and totalling the results (two queries for each university, as above) but to measure inter-university connectivity would take 149x258 = 38,442 queries (there are 258 German University institutions).

Automatic query submission

The major search engine families allow queries to be submitted automatically by programs but only via their official API or web service. These have restrictions in the number of queries that can be submitted. In addition, Yahoo! and Google have the additional limitation that their APIs give results from only a subset of their full database. Hence using MSN has the most potential for good results for link analysis research requiring large numbers of queries.

- Google: 1,000 per user per day (restricted database)
- MSN: 10,000 per computer (Internet Protocol address) per day (full database)
- Yahoo!: 5,000 per user per day (restricted database)

All commercial search engines return up to a fixed maximum number of URLs per query, which generates a methodological problem. Google and Yahoo! return up to 1,000 matching URLs per query, and MSN only 250. This is an issue for link counting research that is ostensibly only interested in the link counts and not individual URLs because previous research has shown that the URL-based Alternative Document Model link counting methods can give better results than simple link (or link page) counting [THELWALL, 2002; THELWALL & HARRIES, 2004]. Although methods have been developed to circumvent the limited number or URLs per query [THELWALL & AL., 2006] these are currently impractical to use on a large scale for multiple queries because

of their need for human intervention and the generation of possibly large numbers of extra queries. Hence the use of standard link page counting is unavoidable for the current study.

Data collection

For the current study and based upon the limitations discussed above we adopted the following procedure. First, we identified the domain names of the universities of the 25 European Union member states (as of January 2006, Bulgaria and Romania became members in January 2007). In addition, we identified the domain names of the universities of 18 further European nations for comparison purposes (see Appendix, Table 1). Using the MSN Web Service and the free LexiURL software (http://lexiurl.wlv.ac.uk/) we submitted queries to MSN over a period of one month to count links (link pages) between each identified university and each identified official European web domain (in both directions). We then consolidated the data by totalling the results for all universities in a country. This gave two matrices. The first matrix recorded the number of links (link pages) from the universities of country i to the country code domain of j. The second matrix recorded the number of links (link pages) to the universities of country j from the country code domain of country i. Note that the MSN link search was withdraw during 2007 after the data had been collected.

Web presence

Figure 1 summarises the four link data sets collected. For convenience we describe the sum of all these statistics as the 'web presence' of a country. Larger web presence values suggest a more active international academic interlinking to and/or from the country and/or its universities. The graph confirms the trend for larger and richer countries to have more links of all kinds associated with them. Figure 1 is difficult to properly interpret, however, precisely because of the varying sizes and wealth of the countries covered.

Figure 2 uses 2004 and 2005 (the latest available for each country) Gross Domestic Product (GDP) data from the World Bank website in order to benchmark links against the wealth of countries [THE WORLD BANK GROUP, 2006]. The approximately linear graph (Pearson correlation 0.905, more reliable non-linear Spearman correlation: 0.942) gives confidence that this comparison is appropriate. We also compared the link data with tertiary education enrolment and population statistics, both of which provided a significantly worse fit. Using a weak benchmarking approach [THELWALL, 2004], the outliers give the most significant information.
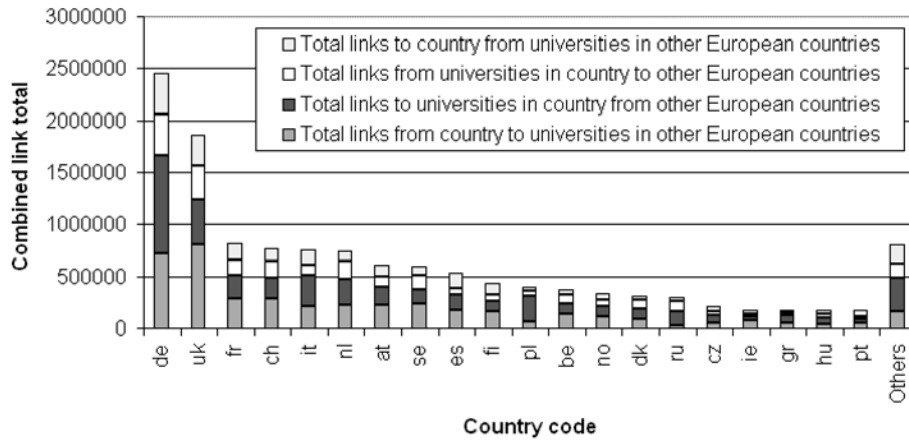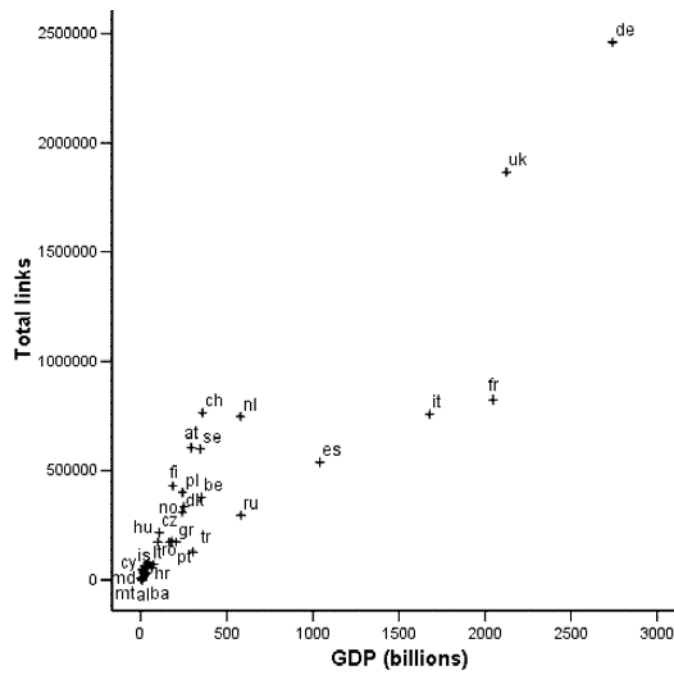
Figure 1. Link pages in the two data sets

Figure 2. Link pages against GDP (US$, 2005 or 2004) for the combined four link data sets

These are countries that attract relatively few/many links for their wealth. The poorest-performing country was Albania (2% of the average links/GDP ratio), with France (29%), Turkey (30%), Bosnia and Herzegovina (32%) and Italy (32%) are following closely. The best-performing countries were Estonia (302%) and Iceland (260%). The best of the larger countries were Austria (148%), Switzerland (153%) and Finland (165%). There was a slight tendency for EU member states to have a higher ratio than the rest, especially the larger countries. There is not, however, a clear tendency for EU countries to be significantly better than the rest.

A possible alternative reading for Figure 2 is that there are three web tendencies at work: high web use (e.g. ch, nl, at, se, fi), medium web use (e.g., de, uk) and low web use (e.g. ru, es, it, fr). These three groups could conceivably be following different patterns although we can only speculate about what the causes could be. One possible cause might be the different structures in education, with countries that have significant and elite government research outside of universities loosing links to the universities as a result. Another possibility is that countries except the UK with a strong international focus outside Europe are disadvantaged (perhaps Spain, France, Portugal), or the same for larger countries that have a strong tradition of using their own language in academic research (e.g. Germany, France, Spain, Russia).

Topology

Total Interlinking (Raw Link Counts)

Visualisations are needed for the strength of interconnectivity between the European nations. Previous similar research has used a variety of different methods. Link diagrams use arrows for links, with arrow widths proportional to link counts and no special positioning of the arrow sources and targets [THELWALL, 2001]. Others have used clustering and visualisation methods designed to identify groups of similar entities rather than the strength of connections. These have used statistical techniques such as hierarchical cluster analysis [MUSGROVE & AL., 2003] or multidimensional scaling (MDS), often with colink counts rather than direct link counts as a similarity measure [ORTEGA & AL., 2007, TO APPEAR; ORTEGA PRIEGO, 2003; POLANCO & AL., 2001; VAUGHAN, 2005; ZUCCALA, 2006]. An alternative choice for a link diagram, and used below, is one with the nodes (countries) positioned with highly interconnected countries close together, and arrow thickness proportional to link counts. The positioning of nodes can be achieved with a standard algorithm such as Kamada–Kawai [KAMADA & KAWAI, 1989], as implemented in the graph drawing program Pajek. We used Kamada–Kawai but adjusted the positions of the nodes to make the diagram more easily interpretable, for example by reducing the number of line crossings and positioning

structurally similar nodes close together to create a visual pattern (e.g. hu, gr, ie, and ru in Figure 3) in line with mathematical graph theory visualization recommendations [BATTISTA & AL., 1999].

From Figures 3 and 4 it can be seen that there is a similar overall pattern of linking although the shapes of the graphs are slightly different. In both cases, and particularly Figure 4, it is clear that there is a core of densely interconnected larger countries plus a periphery of smaller ones that mainly connect to the larger countries. In comparison to a similar previous graph [HEIMERIKS & VAN DEN BESSELAAR, 2006], in Figures 3 and 4, the UK appears more central and interconnected, and the overall network has more structure, in line with the expanded set of countries. The previously observed dominance of the UK and Germany [ORTEGA & AL., 2007] is also apparent but not overwhelming. The importance of Switzerland (ch, not in the EU) is clear from the above map. Poland, a newcomer to the EU, is surprisingly well connected. Perhaps this reflects a strong computing industry or the recent wave of migration. The pattern is of a core and periphery, rather than geo-political linking, although there is an element of both.
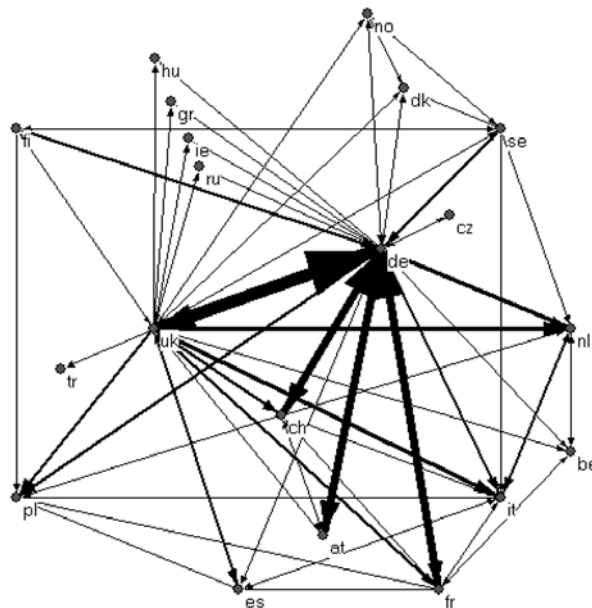

Figure 3. (Links to universities; core and periphery shape) European link network with the width of arrows proportional to the number of pages in the source country domain linking to university web sites in the target country domain. Links below 10% of the maximum are not shown; unlinked countries are not shown
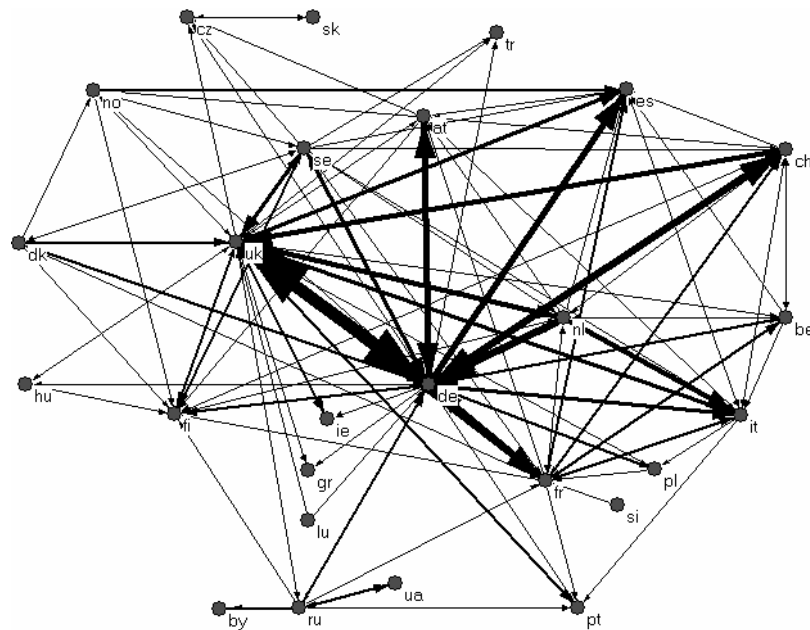
Figure 4. (Links from universities; core and periphery shape) European link network with the width of arrows proportional to the number of pages in the source country universities linking to the target country domain. Links below 10% of the maximum are not shown; unlinked countries are not shown

Normalised network diagrams

Although the network diagrams above show the strongest link connections within the data set, it is also useful to identify pairs of countries with relatively strong ties. For this some kind of size normalisation is needed so that countries with few links can have their main partners identified. For international co-authorship data GLÄNZEL & SCHUBERT (2001) used Salton's measure, the geometric mean. A characteristic is that it is biased towards the countries with most links, despite the normalisation. Network diagrams based upon Salton's measure (not shown) were similar to the non-normalised diagrams (Figures 3 and 4).

An alternative is to use the theoretical approach of the Chi-square test. This assumes that there is no relationship between the countries and calculates an expected value for the links from A to B. The ratio (actual – expected)/expected can be used as an indicator for how far the actual link count deviates from the expected value [LUUKKONEN & AL., 1992; TIJSSEN & MOED, 1989]. Liang, [LIANG & ZHU, 2002] Musgrove [MUSGROVE & AL., 2003] and others have shown that the Chi square contingency table calculations used to calculate the expected values need a correction factor to compensate for the zero

diagonal (i.e., the lack of self-links), and this correction factor is to subtract from the denominator the sum of the target country inlinks. This adjustment removes bias against the larger countries. Note that this adjustment only made a small difference to the diagrams produced with our data set.

From Figure 5 it is clear that after normalisation the strongest links are between the smallest nations and reflect geopolitical considerations. Most of the links above reflect known political connections. For example: the countries of former Yugoslavia are represented and many interlink; similarly for the former countries of the Czech Republic; Greece links to Cyprus, Finland to Estonia. It is possible, however, that this method over-emphasises the connections between small countries because countries with few links may reflect a different linking dynamic [THELWALL & SMITH, 2002].
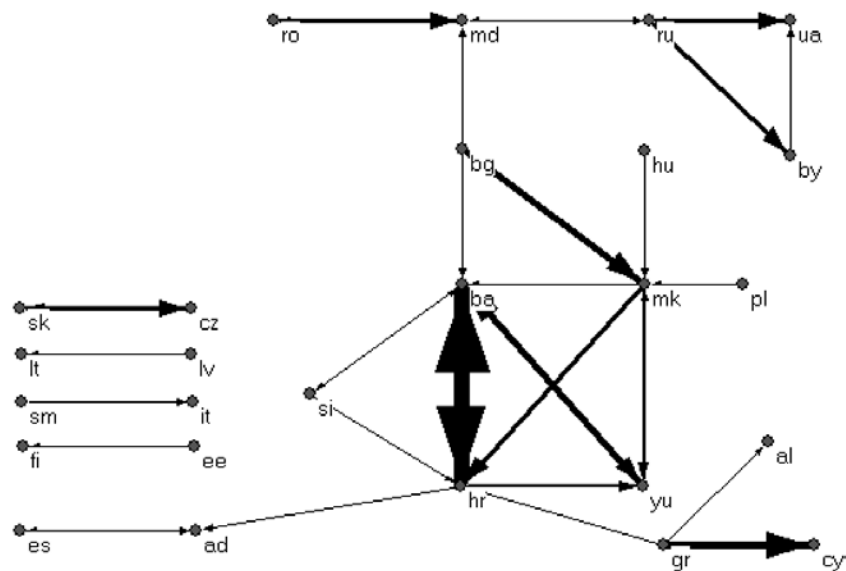

Figure 5. (Fragmented geopolitical map) Strongest links using Liang/Musgrove-corrected Chi-square values. Links below 10% of the maximum are not shown; unlinked countries are not shown

Figure 6 uses the same statistics but reports only the top 13 countries in order to see how they interlink. This normalised graph is again primarily geopolitical. The UK and Germany are present but do not dominate. Interestingly, the UK is connected weakly to many countries but Germany is strongly connected to two other and weakly to three others. All except the UK are neighbours. This suggests that Germany's influence is most strong in its region but the UK may be more widely influential, albeit nowhere strongly influential. Again, Poland has a significant presence as a newcomer.
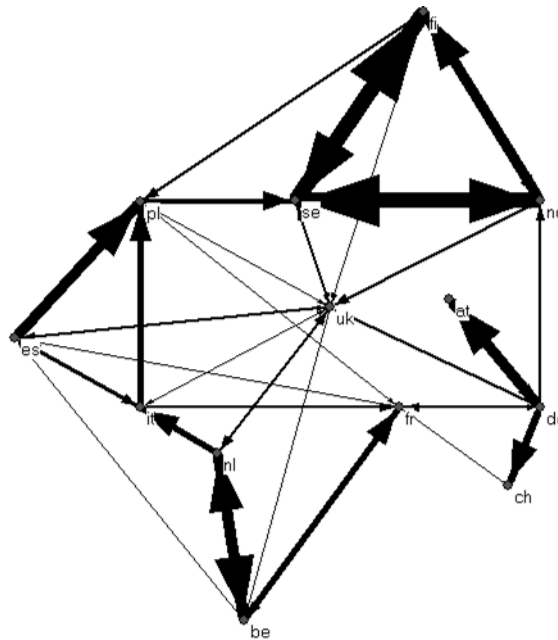
Figure 6. (Connected geopolitical map) Strongest links between the top 13 countries (links to universities) using Musgrove-corrected Chi-square values. Links below 10% of the maximum are not shown

Conclusions

Perhaps the clearest outcome from the analysis is the set of outliers in Figure 2: countries that may be under-represented on the web. Given the importance of the web in academic research, this is an important finding and problem for Europe.

The network diagrams show that the Web, at least as measured by linking, is dominated by the larger Western European nations, primarily the former EU countries but also Switzerland and, perhaps surprisingly, Poland. The web clearly reflects geopolitical influences upon link creation in addition to the influence of the larger, richer countries.

In terms of the methods used, we were unable to get ideal data – that of interlinking between universities – because the scale of the data collection made this impractical with the current generation of web search tools. Hence we had have focused on data about links to individual universities from whole countries and vice versa. Nevertheless, the methods were able to get data that gave believable results. The real value of this research may be seen in the future when comparisons are made in successive years to identify changes in linking patterns, perhaps helping to assess the extent of European

integration. Nevertheless, because of the dependence upon search engines and a wide range of reasons for informal link creation, such conclusions would need to be interpreted cautiously.

References

BARJAK, F. (2005), *NetReAct Deliverable 1.1: Characterisation of Doctoral Students.* Solothurn: University of Applied Sciences Solothurn Northwestern Switzerland.

BATTISTA, G. D., EADES, P., TAMASSIA, R., TOLLIS, I. (1999), *Graph Drawing: Algorithms for the Visualization of Graphs.*New York, NY: Prentice Hall.

FRY, J. (2006), Coordination and control across scientific fields. Implications for a differentiated e-science. In: C. HINE (Ed.), *New Infrastructures for Knowledge Production: Understanding E-science* (pp. 167–188), Hershey PA: Idea Group Inc.

GLÄNZEL, W., SCHUBERT, A. (2001), Double effort = double impact? A critical view at international co-authorship in chemistry. *Scientometrics,* 50 (2) : 199–214.

HEIMERIKS, G., HÖRLESBERGER, M., VAN DEN BESSELAAR, P. (2003), Mapping communication and collaboration in heterogeneous research networks. *Scientometrics,* 58 (2) : 391–413.

HEIMERIKS, G., VAN DEN BESSELAAR, P. (2006), Analyzing hyperlink networks: The meaning of hyperlink-based indicators of knowledge. *Cybermetrics,* 10 (1), Retrieved August 1, 2006 from: http://www.cindoc.csic.es/cybermetrics/articles/v2010i2001p2001.html.

INGWERSEN, P. (1998), The calculation of Web Impact Factors. *Journal of Documentation,* 54 (2) : 236–243.

KAMADA, T., KAWAI, S. (1989), An algorithm for drawing general undirected graphs. *Information Processing Letters,* 31 (1) : 7–15.

LI, X., THELWALL, M., WILKINSON, D., MUSGROVE, P. B. (2005A), National and international university departmental web site interlinking, part 1: Validation of departmental link analysis. *Scientometrics,* 64 (2) : 151–185.

LI, X., THELWALL, M., WILKINSON, D., MUSGROVE, P. B. (2005B), National and international university departmental web site interlinking, part 2: Link patterns. *Scientometrics,* 64 (2) : 187–208.

LIANG, L., ZHU, L. (2002), Major factors affecting China's inter-regional research collaboration: Regional scientific productivity and geographic proximity. *Scientometrics,* 55 (2) : 287–316.

LUUKKONEN, T., PERSSON, O., SIVERTSEN, G. (1992), Understanding patterns of international scientific collaboration. *Science, Technology, & Human Values,* 17 (1) : 101–126.

MAYR, P., TOSQUES, F. (2005), *Google Web APIs: An Instrument for Webometric Analyses?* Retrieved January 20, 2006, from http://www.ib.hu-berlin.de/%7Emayr/arbeiten/ISSI2005_Mayr_Toques.pdf

MIDDLETON, I., MCCONNELL, M., DAVIDSON, G. (1999), Presenting a model for the structure and content of a university World Wide Web site. *Journal of Information Science,* 25 (3) : 219–227.

MUSGROVE, P. B., BINNS, R., PAGE-KENNEDY, T., THELWALL, M. (2003), A method for identifying clusters in sets of interlinking Web spaces. *Scientometrics,* 58 (3) : 657–672.

ORTEGA, J. L., AGUILLO, I., COTHEY, V., SCHARNHORST, A. (2008), Maps of the academic web in the European Higher Education Area: an exploration of visual web indicators. *Scientometrics,* 74 (2) : 295–308.

ORTEGA PRIEGO, J. L. (2003), A Vector Space Model as a methodological approach to the Triple Helix dimensionality: A comparative study of Biology and Biomedicine centres of two European national research councils from a Webometric view. *Scientometrics,* 58 (2) : 429–443.

PARK, H. W., THELWALL, M. (2006), Web science communication in the age of globalization: Links among universities' websites in Asia and Europe. *New Media and Society,* 8 (4) : 631–652.

POLANCO, X., BOUDOURIDES, M. A., BESAGNI, D., ROCHE, I. (2001), *Clustering and Mapping Web Sites for Displaying Implicit Associations and Visualising Networks*: University of Patras.

THE WORLD BANK GROUP. (2006), *EDSTATS - Global Country Data*. Retrieved July 30, 2006, from http://devdata.worldbank.org/edstats/cd5.asp

THELWALL, M. (2001), Exploring the link structure of the Web with network diagrams. *Journal of Information Science,* 27 (6) : 393–402.

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

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

THELWALL, M. (2004), Weak benchmarking indicators for formative and semi-evaluative assessment of research. *Research Evaluation,* 13 (1) : 63–68.

THELWALL, M., HARRIES, G. (2004), Do the Web sites of higher rated scholars have significantly more online impact? *Journal of the American Society for Information Science and Technology,* 55 (2) : 149–159.

THELWALL, M., LI, X. (2005), *Web Networks in Life Sciences Research (NetReAct D2.4),* Bonn, Germany: Empirica Gesellschaft für Kommunikations- und Technologieforschung mbH.

THELWALL, M., SMITH, A. G. (2002), A study of interlinking between Asia-Pacific university web sites. *Scientometrics,* 55 (3) : 335–348.

THELWALL, M., TANG, R., PRICE, E. (2003), Linguistic patterns of academic web use in Western Europe. *Scientometrics,* 56 (3) : 417–432.

THELWALL, M., VANN, K., FAIRCLOUGH, R. (2006), Web issue analysis: An Integrated Water Resource Management case study. *Journal of the American Society for Information Science & Technology,* 57 (10) : 1303–1314.

THELWALL, M., VAUGHAN, L., BJÖRNEBORN, L. (2005), Webometrics. *Annual Review of Information Science and Technology,* 39 : 81–135.

TIJSSEN, R., MOED, H. F. (1989), Literature-based statistical analyses of international scientific co-operation, an exploratory case study of the Netherlands. In: A. F. J. VAN RAAN, A. J. NEDERHOF, H. F. MOED (Eds.), *Science Studies (Select Proceedings of the First International Workshop on Science and Technology Indicators, Leiden, 1988)* (pp. 129–146), Leiden: DSWO-Press.

VAUGHAN, L. (2005), Exploring website features for business information. *Scientometrics,* 61 (3) : 467–477.

VAUGHAN, L. (2006), Visualizing linguistic and cultural differences using Web co-link data. *Journal of the American Society for Information Science and Technology,* 57 (9) : 1178–1193.

VAUGHAN, L., THELWALL, M. (2005), A modeling approach to uncover hyperlink patterns: The case of Canadian universities. *Information Processing & Management,* 41 (2) : 347–359.

ZUCCALA, A. (2006), Author cocitation analysis is to intellectual structure as web colink analysis is to…? *Journal of the American Society for Information Science & Technology,* 57 (11) : 1487–1502.

Appendix

Table 1. Country codes and country names. EU countries are in bold,
and those that joined on 1 May 2004 are italicised.
Bulgaria and Romania joined in January 2007

Country code	Country
at	**Austria**
be	**Belgium**
cy	*Cyprus*
cz	*Czech Republic*
de	**Germany**
dk	**Denmark**
ee	*Estonia*
es	**Spain**
fi	**Finland**
fr	**France**
gr	**Greece**
hu	*Hungary*
ie	**Ireland**
it	**Italy**
lt	*Lithuania*
lu	**Luxembourg**
lv	*Latvia*
mt	*Malta*
nl	**Netherlands**
pl	*Poland*
pt	**Portugal**
se	**Sweden**
si	*Slovenia*
sk	*Slovakia*
uk	**UK**
ad	Andorra
al	Albania
ba	Bosnia and Herzegovina
bg	*Bulgaria*
by	Belarus
ch	Switzerland
hr	Croatia
is	Iceland
li	Liechtenstein
md	Moldova
mk	Macedonia
no	Norway
ro	*Romania*
ru	Russia
sm	San Marino
tr	Turkey
ua	Ukraine
yu	Serbia and Montenegro