

A statistical analysis of the web presences of European life sciences research teams

Franz Barjak, *School of Business, University of Applied Sciences Northwestern Switzerland, Riggbachstrasse 16, CH-4600 Olten, Switzerland. E-mail: franz.barjak@fhnw.ch Tel: +41 62 287 7825 Fax: +41 62 287 7845*

Mike Thelwall, *School of Computing and Information Technology, University of Wolverhampton, Wulfruna Street, Wolverhampton WV1 1SB, UK. E-mail: m.thelwall@wlv.ac.uk Tel: +44 1902 321470 Fax: +44 1902 321478*

Abstract

Web links have been used for around ten years to explore the online impact of academic information and information producers. Nevertheless, few studies have attempted to relate link counts to relevant offline attributes of the owners of the targeted web sites, with the exception of research productivity. This paper reports the results of a study to relate site inlink counts to relevant owner characteristics for over 400 European life science research group web sites. The analysis confirmed that research group size and web presence size were important for attracting web links, although not research productivity. Little evidence was found for significant influence of any of an array of factors, including research group leader gender and industry connections. In addition, the choice of search engine for link data created a surprising international difference in the results, with Google perhaps giving unreliable results. Overall, the data collection, statistical analysis and results interpretation were all complex and it seems that we still need to know more about search engines, hyperlinks, and their function in science before we can draw conclusions on their usefulness and role in the canon of S&T indicators.

Introduction

In recent years a growing body of research has used hyperlinks to generate webometric indicators. In particular, the number of hyperlinks that point to a web document from other Internet documents has been conceived as an indicator of the impact of this document and its producer(s) on the internet (Ingwersen, 1998). A high 'web impact' or 'online impact' for a document often signals that it might contain information that may be useful for visitors to the source documents of the links, although links are also useful for other purposes such as acknowledgement.

Most studies of the web impact of academic web sites have been carried out for datasets of university or departmental web sites. They have assessed the relationship between web impact and other impact measures: between hyperlinks to organizations and their research performance measured through peer-review ratings or publication impact (Li, Thelwall, Musgrove, & Wilkinson, 2003; Tang & Thelwall, 2003; Thelwall, 2001). Further control variables like the scientific discipline, and country have been included in these studies. Few explorations of the links to scholars' individual homepages have appeared, taking into account homepage content, personal characteristics, and institutional characteristics of the homepage owners (Barjak, Li, & Thelwall, in press; Nelson, 2005).

Though these analyses shed some light on hyperlinking practices in the academic web and the relationship between web linking and some offline variables, further studies are needed to evaluate the potential of hyperlinks as a basis for valid and reliable S&T indicators. For this purpose the meaning of hyperlinks and their potential of representing other concepts in science but outside of cyberspace, such as research performance, collaboration activities, or scientific networks at the level of countries, universities, parts of universities or even individual scholars needs to be clarified. Then the World Wide Web could be added as an additional database for the description and analysis of scientific processes and structures.

In this paper we use a variety of statistical methods in an attempt to explain why research groups attract hyperlinks. It deals with two shortcomings which have not yet been discussed sufficiently in the available literature:

1. Most previous papers use hyperlinks in the academic web collected by means of one search engine only, or a single web crawler. Our paper compares hyperlink data collected through three different current major search engines, namely Google, Yahoo!, and MSN.
2. In a previous paper we analysed the factors explaining the web impact of scientists' personal homepages (Barjak, Li, & Thelwall, in press). This paper goes beyond homepages and uses the entire web presence of a team: homepages plus additional pages.

The article uses different methods of correlation and regression analyses and follows what has been called an indirect approach for interpreting the role of links (Thelwall, 2006). The main weakness of this approach is, without doubt, the danger of spurious correlations and the omission of underlying factors. The risk of spurious correlations can be reduced if the selection of variables for such a correlation analysis is informed by a theoretical framework or

model. In addition, the latter should provide reasons why the variables included in a correlation analysis should affect the number of links in the chosen research population and setting. However, such a model has not yet been developed. Based on the results of previous studies we will make a first attempt in the next section. The model and justification can also help to reduce the risk that important underlying factors have been omitted.

Theoretical background

The literature on hyperlinks in the academic web and their relationship to the information that is represented on web sites, the characteristics of the web page owners, or the environment in which they work (e.g. academic discipline, country) is increasing. Without going into detail, we can summarise a number of stylised facts that result from this work:

Information-related factors

1. Many hyperlinks emanate from hubs (lists) and point to targets which carry actual information (“hub-authority nature of the web”) (Bar-Ilan, 2005; Kleinberg, 1999).
2. A large percentage of links are related to scholarly activity (Bar-Ilan, 2004; Wilkinson, Harries, Thelwall, & Price, 2003).
3. The quality of link data is reduced notably by “noise” or anomalies (Thelwall, 2003; Thelwall, 2004; Thelwall & Harries, 2004).
4. The more information is included in a web document, the more links point to it (Li, Thelwall, Wilkinson, & Musgrove, 2005; Thelwall & Harries, 2004).
5. The higher the quality (e.g. usefulness, credibility) of the information in a web document, the more links point to it (Barjak, Li, & Thelwall, in press; Fogg et al., 2001; Liu, 2004; Park, Barnett, & Nam, 2002).
6. Quality is not an absolute measure but relative, depending for instance on the timeliness of the information (Beaulieu & Simakova, 2006) or the availability of information on other sites.
7. The longer a web page has maintained the same URL, the more inlinks it receives (“preferential attachment”) (Adamic & Huberman, 2000; cf. Barabási & Albert, 1999).

Owner-related factors

8. The higher the peer recognition of the page-owners, the higher the number of links to their web documents:

- at the level of universities (Smith & Thelwall, 2002, 2005; Thelwall, 2001, 2002a; Thelwall & Harries, 2003, 2004; Thelwall & Wilkinson, 2003),
 - and departments (Chen, Newman, Newman, & Rada, 1998; Fry, 2006; Li, Thelwall, Musgrove, & Wilkinson, 2003; Li, Thelwall, Wilkinson, & Musgrove, 2005),
 - but only sporadically at lower levels: individual scientists (Barjak, Li, & Thelwall, in press).
9. The higher the impact of the work of the page-owners (departmental level), the higher the number of links to their web presences (Li, Thelwall, Wilkinson, & Musgrove, 2005; Tang & Thelwall, 2003).
 10. Higher rated scholars do not produce web content with higher average impact, but they produce more content and hence gain a higher total impact (Thelwall & Harries, 2004).
 11. Only few hyperlinks represent communication or collaboration relations between the owners (Bar-Ilan, 2004; Beaulieu & Simakova, 2006; Fry, 2006; Heimeriks, Hörlesberger, & Van den Besselaar, 2003; Thelwall, 2003).
 12. Hence, no clear relationship between the amount of collaboration and inlinks to the web presences is measurable (Barjak, Li, & Thelwall, in press; Heimeriks & Van den Besselaar, 2004).
 13. Any influence of page owners' demographic characteristics on inlinks is not confirmed (Barjak, Li, & Thelwall, in press; Thelwall, Barjak, & Kretschmer, 2006).

Environment-related factors

14. The number of links differs between scientific disciplines and research fields (Barjak, Li, & Thelwall, in press; Li, Thelwall, Wilkinson, & Musgrove, 2005; Nelson, 2005; Tang & Thelwall, 2003, 2004; Thelwall, Harries, & Wilkinson, 2003; Thelwall, Vaughan, Cothey, Li, & Smith, 2003).
15. Link patterns (e.g. sources of inlinks) differ between disciplines (Harries, Wilkinson, Price, Fairclough, & Thelwall, 2004; Thelwall, 2004).
16. The number of links differs between countries (Barjak, Li, & Thelwall, in press; Li, Thelwall, Wilkinson, & Musgrove, 2005; Thelwall, Vaughan, Cothey, Li, & Smith, 2003).
17. The larger the distance between two universities the fewer the number of links that connects them (Heimeriks & Van den Besselaar, 2006; Thelwall, 2002b, 2002c).

Integrating these findings into a model we get the representation shown in Figure 1. The straight lines stand for relationships which have been found in previous empirical work and the dotted lines stand for relationships which are still tentative and not confirmed. The + and – indicate the direction of the relationship. On the left, the model includes factors known from non-web research showing a negative effect of distance on research communication and collaboration (Allen, 1991; Cummings & Kiesler, 2003; Kraut, Egido, & Galegher, 1990); the role of demographic characteristics on scientists resource access and success has been discussed in the gendered science literature (Etzkowitz, Kemelgor, Neuschatz, & Uzzi, 1992; Prpic, 2002; Shauman & Xie, 2003); and the importance of issues such as field, organisation, country, summarised here as “environmental issues”, for scientific practice and success is evident in a broad range of research (see e.g. the overview in (Barjak, 2006).

Comment: Add Katz 1994?

Figure 1: A model for explaining the number of external site inlinks to web presences in the academic web

[Figure 1 about here]

The empirical work in this paper will investigate the relationships represented in the model further at the level of research teams, and in particular the following questions.

- Is there a direct relationship between research productivity, impact, and recognition of the web page owners and the external site inlinks to their web presences?
- Do research groups which collaborate a lot receive more inlinks, i.e. do inlinks reflect real world collaboration?
- Is there a gender gap in site inlinks?

Search engines in webometric research

Search engines have been used to collect data for webometrics at least since AltaVista first launched its link search capability (Ingwersen, 1998). Each of the three current major families of search engines (Google, Yahoo!, MSN) provides a facility (interface) for programs to automatically submit queries (e.g., Mayr & Tosques, 2005). This is a big advantage for Webometrics research involving thousands of searches. For Google and Yahoo!, however, the results seem to reflect only a part of the results available from the main search page. A

significant body of research has also analysed the reliability of search engine results (Bar-Ilan, 2004).

Commercial search engines have probably never found all existing web pages. For example, in 1999 the largest coverage was estimated to be that of Northern Light, at 16% (Lawrence & Giles, 1999). Since then, the introduction of virtual servers and the extensive use of dynamic pages has made it impossible to sensibly estimate the proportion of the web that modern search engines cover (Thelwall, 2002d). Nevertheless, something can be said about the overlaps between the coverage of different engines. Since search engines index the web mainly by following links, it seems likely that all will cover most of a central core of interconnected pages (Broder et al., 2000). Nevertheless, there was a surprisingly low degree of overlap, at least in 1999 (Lawrence & Giles, 1999). Differences in coverage may be accounted for by a number of reasons, including different sets of new URLs submitted by users, different policies for identifying and removing duplicate and spam pages, differing coverage of non-HTML file types, and the extent of dynamic page indexing. In addition there may be differing upper limits to the number of pages to index per site, the maximum size of a page to index, and total database size. These factors are commercial secrets but some research has evaluated the results of a range of searches, leading to a number of implications.

First, search engine results are normally estimates (e.g., “about 50,000 matches”) and in the past have been occasionally wrong by several orders of magnitude (Bar-Ilan, 1999; Rousseau, 1999; Snyder & Rosenbaum, 1999). The estimates may be extrapolations made after processing only a portion of the database, hence the possibility of errors. Accuracy can sometimes be improved by checking the second (or last) page of results for its estimate. Second, systematic biases in coverage are present with older and better-linked-to sites probably over-represented (Lawrence & Giles, 1999; Vaughan & Thelwall, 2004). Third, results may not be internally consistent (Snyder & Rosenbaum, 1999). For example identical searches at the same time may give different answers (Mettrop & Nieuwenhuysen, 2001) and narrowing down a search may occasionally increase the number of matches. Fourth, search engines have undocumented and obscurely-documented features, particularly for infrequently-used types of query such as link searches (e.g., “Collecting the web data” below). There may also be undocumented informal limits such as the maximum number of terms or characters per query.

Data and Methods

The data for the paper was generated through different methods of data collection for a sample of research teams in the life sciences defined according to the ISCED 1997 category 42: an on-line survey, the retrieval of bibliometric data from the ISI Thomson database, and last but not least webometric data collected from search engines. The overall dataset contains more than 400 research teams and covers 10 European countries.

Collecting the web data

The list of research group URLs was submitted to Google, MSN and Yahoo! via their programming interfaces to estimate three relevant statistics.

1. The number of pages in the web site, as covered by search engines.
2. The total number of links to the web site or web page.
3. The number of links to the web site from other domain names (site inlinks).

The advanced query keywords below were used as the basis for the data gathering.

- **inurl:** reports the number of pages with URLs matching the command.
- **link:** reports the number of pages that link to the *exact* specified URL (Google) or that have links *containing* the specified URL (MSN, Yahoo!)
- **linkdomain:** reports the number of pages that link to any pages in the specified domain.
- **site:** reports the number of pages with domain names equal to, or ending in, the specified domain name (or part of domain name).

The above commands can be combined together by using Boolean operators like AND and NOT (or -). For example, `linkdomain:x -site:x` matches pages with a link to any page in site x, but excluding all pages in site x (i.e., site inlinks).

After extensive checking of results and repeated attempts at constructing accurate queries, we devised a set of procedures to construct the queries to gather the three relevant types of statistic. Different queries were used for each search engine and for each type of URL, depending upon whether the research group possessed its own domain name or not. The objective was to obtain as complete and accurate information as possible, and using human intervention where necessary. Table 1 summarises the queries, using the following codes.

- u is the full URL of research group, except the initial `http://`

- d is the domain name of URL
- p is the path of URL (following the domain name and slash)
 - pm is p except with slashes inserted between numbers and letters, whenever adjacent.
 - pg is p path except with ampersands replaced by slashes.
 - py is p path except with any of html, htm, asp, cgi-bin, doc, php, index, id, org, shtml, dep removed.

To illustrate the codes, the site <http://www.dundee.ac.uk/biocentre/SLSBDIV1tks.htm> gives: u=www.dundee.ac.uk/biocentre/SLSBDIV1tks.htm, d= www.dundee.ac.uk and p=biocentre/SLSBDIV1tks.htm. Now pg is the same as p, but pm=biocentre/SLSBDIV/1/tks.htm and py = biocentre/SLSBDIV/1/tks.

The three different path variables py, pg and pm were needed because the inurl: query seems to be used by all search engines as an additional search term and so has some peculiarities that are targeted at users who are looking for page content rather than specific URLs. For example, all the search engines split the URL into separate “words”, and match the URLs against the words in any order. MSN also splits alphabetic and numeric characters into separate words. For example inurl:x2 would not match an URL containing x2 but inurl:x/2 would because x and 2 would match separately. Some of the anomalies in the inurl: command could not be circumvented. For example, MSN sometimes split alphabetical words into two or more separate words, and sometimes appeared not to index a word in the URL. For example MSN’s inurl: ignores the pseudo-word nanobiotech. In such cases we submitted more general searches to assess how many URLs were indexed by MSN. Yahoo!’s inurl: ignored a set of common words so we manually identified and removed these from the query and then manually checked the results to ensure that the simplified query had not allowed any incorrect matches to be returned.

Table 1. Queries submitted to search engines – for Google the inlink commands match links to the exact URL, whereas MSN and Yahoo! inlink commands may match multiple URLs.

[Table 1 about here]

We constructed two web statistics for the data analysis. The first is site size, as indexed by search engines. We defined this to be the log of one plus the maximum number of pages found in the site by any search engine. One was added to avoid taking the log of zero. We used the log of site size because web data is known to follow a power law. Search engines do not index the whole web and so taking the maximum result is a reasonable method to estimate the size of the indexable part of a web site. This is only an approximation, however, since other search engines could give more complete coverage and because the numbers reported by search engines can be estimates, particularly for large sites.

Similarly for the inlink count statistics, we used the log of one plus the maximum number of site inlinks found to the site by any search engine. For this, the figures from Google should logically be ignored since it counts only links to the specified page and reports a sample of its results and not its full results. Nevertheless, it was used because for small sites it occasionally found links that the other search engines had not found.

Data analysis

The data analysis makes use of different statistical methods. In addition to bivariate analyses we employ multivariate count data models relating hyperlinks to structural characteristics of the research teams (team size, team age, team composition, country, sub-discipline), their research performance (number of publications, citations, relative impact), and collaboration variables (overall number of collaborating groups and some structural variables).

The nominal explanatory variables in the dataset – country, academic discipline, type of affiliation, gender, level of recognition – were included as [0, 1] coded dummy variables, e.g. the “Czech Republic” variable has the value “1” for all Czech teams and “0” for all teams from other countries. As is standard econometric practice, one of the variables in each group was left out. This variable is the reference category (and expressed in the value of the constant together with the reference categories of the other dummy variables). For instance, among the country variables, the variable identifying teams from Sweden was excluded from the estimations and the values of the country dummy variables have to be interpreted in relation to the Swedish teams.

Note that the independence of the underlying data is reduced by the phenomenon of copying in citations (the “Matthew effect”) and web links (the “rich get richer” phenomenon). This is important because part of the reason why a web site attracts links may be that it already has links, rather than because of intrinsic properties (e.g. research ability) of the page owner (Barabási & Albert, 1999; Pennock, Flake, Lawrence, Glover, & Giles, 2002).

The analysis firstly used the raw link counts in order to find out whether differences between the search engines appear. For the multivariate models of the raw link counts we used count data models. The baseline approach of count data models is a Poisson regression model which better accounts for nonnegative and integral data than for instance the ordinary least squares regression model. If the dependent variable is subject to overdispersion – the variance exceeds the mean – the negative binomial regression model (NEGBIN) is preferable, as it permits this difference (Cameron & Trivedi, 1998). We tested for overdispersion as described in Cameron and Trivedi (1998) and include the alpha values from the NEGBIN estimation in the results tables – significant alphas indicate overdispersion. Moreover, if the dataset contains many zeros (“zero inflated” or ZI) either Zero Inflated models or Hurdle models can deal with this. The Vuong statistic is proposed as a test statistic for zero inflation. It is distributed as standard normal with a critical value of 1.96, i.e. a value of more than +1.96 favours and less than -1.96 rejects the ZI NEGBIN model (Greene, 2000) According to the results of the Vuong Statistic we estimated and presented ZI NEGBIN models.

Moreover, the analysis uses the logged external site inlinks per team in Ordinary Least Squares (OLS) regressions. Through logged data the problem of not normally distributed residuals, a precondition for OLS estimations, can be reduced. We calculated the Bowman and Shenton chi square test statistic (BS chi), a test that is based on the skewness and kurtosis of the variable, to control for the normality of the residuals (as proposed in Econometric Software, 2002 and implemented in LIMDEP 8.0). The test results led us to reject the assumption of normally distributed residuals. Upon closer inspection two problems became apparent which explain the test results:

- (1) The large share of research teams with 0 or 1 external site inlinks. We estimated Tobit models for censored data which can deal with an overrepresentation of zeros – however, this did not solve the normality problem as the CM test statistic (CM chi) of the Tobit models shows.

- (2) The heteroscedasticity of the residuals, i.e. the value of the residuals depends on the explanatory variables of the model – in our case in particular on the page count. The Breusch-Pagan test (BP) and the likelihood ratio test (LR) indeed show that heteroscedasticity is a problem in the estimations. The White estimator and Weighted Least Squares Regressions (WLS) are suggested as possible solutions for heteroscedastic residuals and they were implemented in the estimations (Econometric Software, 2002).

The time lags between the different data sets for research teams are a weakness of the current paper: the survey was carried out in 2005 but it asked for information on 2003; the bibliographic data retrieved from the Web of Science refers to the year 2001 and the link figures were collected in 2005. This certainly affects the results to some extent and we cannot tell exactly how. However, we nevertheless believe in the validity of our results. Firstly, the survey variables included in the analysis are structural variables, which change slowly. Second, we presume that real world structures and events are reflected in hyperlink creation. For instance, a team who publishes a lot (in journals, other print media and the internet) should receive hyperlinks due to this productivity, but not the other way around. Therefore, it is less problematic that the publication data refers to an earlier year than the hyperlink data. If we take into account a time lag between publication and link creation, much like in citation studies, a time lag in the data is actually recommendable. Whether this lag should be 4 years or not would need time series data, which is not available in our case.

Comparisons of results between Google, Yahoo! and MSN

The search engines returned quite similar results except that Google reported a low number of links, and a high number of pages, as Table 2 shows. Although there were significant differences between the search engines in the *total* number of links and pages found, this tended to be caused by a small number of sites with many pages or links, and the results tended to be broadly consistent. There was a high Spearman correlation between the different search engines (Table 3) which shows that the rank order of the link counts per research group is similar across the different search engines and confirms this consistency. The values in bold indicate similar measured quantities. They are all high, suggesting that for the purposes for which we are using search engines, the results of each could be similar to the results of the others. In other words the choice of search engine may not greatly affect the results of the bivariate and multivariate analyses of the sections below but this is not guaranteed because the two analyses use the raw data rather than the rank orders and because the correlations are significantly short of 1.

Table 2. Means and medians of the search engine results.

[Table 2 about here]

Table 3. Spearman correlations between results (all significant at $p=0.001$).

[Table 3 about here]

Results of the bivariate analyses

Inlinks, nationality, web content and group size

We first look at the bivariate relationships between the site inlink indicators and the structural characteristics of the life science teams. We limit the data to the most significant results; further results can be obtained from the authors upon request.

The cross-country analysis of the site inlink data clearly shows significant differences even in this set of countries with rather developed life science research systems. The raw link data, in particular the data from Yahoo! and MSN, tends to exaggerate the differences somewhat as outliers affect the mean values. Columns 5-8 in Table 4 are therefore better suited for comparison. German teams have significantly more pages than the 10-country average ($101.53 - 1 = 33$ versus $100.97 - 1 = 8.3$), whereas Spanish, Italian, and Norwegian teams have fewer pages. German and UK teams received more than average external inlinks from other websites, whereas Spanish, Italian, and Norwegian teams received fewer. As a next step we normalised the logarithmic link score by subtracting the logarithmic page score which is equal to taking the log of the quotient of links and pages. For this link per page indicator, the leading position of Germany is inverted and it has the fewest links (0.24). In the UK, the country with the highest value, we counted 0.76 links per page. If we employ a different normalisation and use the scientific staff instead of the web pages per team, the leading position of Germany is restored (1.0 links per scientist) whereas the lowest link scores per scientific staff member are found in Spain, Italy, Portugal, France, and the Czech Republic. Two interpretations of this finding are possible: assuming that the content per page is more or less similar across all countries, we would conclude that German teams (and Hungarian teams which are somewhat similar) receive more inlinks because they produce more content (and more pages). If, however, German and Hungarian teams distribute the same amount of content as the other teams on more pages, for instance because they employ a different approach to web publishing, then the page-normalised link count might produce a wrong impression. We do not have any evidence to support either explanation.

Table 4: Site inlinks of teams by country (arithmetic mean with standard error in brackets)

[Table 4 about here]

a Logged value of the maximum number of pages (site inlinks) found by any of the three search engines Google, Yahoo!, or MSN.

b Logged maximum site inlinks – Logged maximum page count

c Logged maximum site inlinks – Logged total scientific staff of the team
d F-statistic: ANOVA procedure; Chi-square: Kruskal-Wallis-Test
significance levels ** = 0.01, * = 0.05, + = 0.1.
Source: NetReAct, authors.

Correlating the raw link counts with team size we get low but significant positive Spearman correlation coefficients: 0.17** for Google links, 0.20** for Yahoo! links and 0.19** for MSN links (stars indicate significance levels, see notes to Tables). The larger a team the more pages it produces too, as the correlation coefficients for the logged page count shows (Pearson/Spearman: 0.12*/0.17**). No correlation between team size and links per page is visible, but team size and links per scientific team member are negatively correlated (Pearson/Spearman: -0.20**/-0.24**). The larger a team, the more web pages it produces and the more links it receives, but for one additional team member *less* than one additional link share is obtained. This would be consistent with teams getting some links just for existing (e.g., from exhaustive link lists of all research groups within a field) and some for their activities (Pennock, Flake, Lawrence, Glover, & Giles, 2002).

Publications, Citations and Inlinks

Next we evaluated the relationship between publications, citations, and site inlinks. The publication data covers articles listed in the *Science Citation Index Expanded (SCIE)* in 2001 and citations are the citations to these articles in the years 2002-2004. The relationship between the bibliographic variables and between hyperlinks is rather weak (see Table 5). Only for the total publication output per team do we find a stable and significant relationship to the hyperlink data and the size of the web presences (logged page count). This relationship might be caused by team size, because the bigger a team the more output it produces, both formal output listed in *SCIE* and informal output that can be put on the web. However, we also obtain a very small but significant correlation coefficient between staff-normalised publication and site inlink counts – this could point to an additional relationship that goes beyond mere size effects. Citations are only weakly related to the logged web page count result but not to site inlinks.

Table 5: Correlations between different variables for publications and citations and site inlinks (Pearson/Spearman correlation coefficients)

[Table 5 about here]

a Log site inlinks (max.) – Log page count (max.)
b Log site inlinks (max.) – Log total scientific staff of the team

significance levels ** = 0.01, * = 0.05, + = 0.1.

Source: NetReAct, authors.

Collaboration

Another issue included in the analysis concerns (research) collaborations. The teams' collaborations were measured in a number of different ways:

- The team leaders were asked in the survey with how many other teams they collaborated,
- from the *SCIE* publication data the numbers of co-authored papers per team were counted, and
- the survey included a question on the funding of doctoral students and post-docs through projects with industry.

From this data we developed a variety of indicators for total collaboration, collaboration with international partners and collaboration with industry. The relationships between the webometric data and the collaboration data are shown in Table 6.

The most remarkable results are similar to the results for the bibliographic variables: the total amount of collaboration partners (number of teams, publications with other/foreign teams) is weakly related to the number of site inlinks, and the staff-normalised collaboration data is even more weakly related to the site inlinks per staff member. In addition, we see a more pronounced relationship for international collaborations (Table 6). These results are consistent with collaboration helping to generate inlinks to a site either directly to acknowledge collaboration (e.g., Harries, Wilkinson, Price, Fairclough, & Thelwall, 2004) or indirectly as a result of increased visibility (e.g., Rousseau, 1992). Teams co-authoring publications with industry partners (based on 2001 bibliographic data) also received slightly more site inlinks ($10^{0.86} = 7.3$) compared to teams without co-authors from industry ($10^{0.63} = 4.3$). This difference is significant in ANOVA and Kruskal-Wallis-Tests at $p = 0.05$. University-industry relationships which were accompanied by industry funding for PhD students or post-docs had no visible influence on inlinks and the results are not shown.

Table 6: Correlations between different variables for collaborations and site inlinks (Pearson/Spearman correlation coefficients)

[Table 6 about here]

a Log site inlinks (max.) – Log page count (max.)
b Log site inlinks (max.) – Log total scientific staff of the team
significance levels ** = 0.01, * = 0.05, + = 0.1.
Source: NetReAct, authors.

Gender and recognition

In addition to the team variables, we also included two variables for the team leaders, namely their gender and their level of professional recognition. Looking at gender differences (Table 7), we see that the raw link counts with Google and MSN differ significantly between male and female team leaders, whereas the data from Yahoo! and the logged data do not differ significantly. If we factor out the size of the web presences or the research teams and take the normalised data, we see that the differences between inlinks to teams with male and female team leaders are probably not due to the gender of the team leader, but the sizes of the web presences and teams: Both the page-normalised and the staff-normalised link counts have virtually identical averages for male and female led teams.

Table 7: Site inlinks by gender of the team leader^a

[Table 7 about here]

a Arithmetic mean (standard error in brackets)
b Log site inlinks (max.) – Log page count (max.)
c Log site inlinks (max.) – Log total scientific staff of the team
d F-statistic: ANOVA procedure; Chi-square: Kruskal-Wallis-Test
significance levels ** = 0.01, * = 0.05, + = 0.1.
Source: NetReAct, authors.

Another criterion for distinguishing the inlink scores to the teams' web pages was the professional recognition of the team leader. This recognition was measured through the answers to a five-item question asking about awards, organisation of international conferences, service on professional committees, editorial boards, and advisory committees within the previous five years. The more these services were rendered, the higher the assessed level of recognition. A nearly significant relationship between the level of recognition and site inlinks appeared only for the MSN inlink data and it was only significant at the hardly acceptable level of $p = 0.1$. The same applies to the logged page count. For the links retrieved with Google and Yahoo!, as well as the page and staff normalised logged maximum links, no significant differences were found.

In addition to the displayed results, we ran the bivariate analyses only with the set of countries with the highest site inlink scores, namely Germany, France, Hungary, Sweden, and the UK. The results do not differ significantly except for the very last variable displayed in Table 10:

the industrial funding of PhD students or post-docs. Teams without this type of funding received significantly more site inlinks overall (6.3) and normalised per scientist (1.4) than teams with this type of funding relationship (4.3 and per scientist: 0.3). This finding is from a smaller set of about 260 teams only and we do not want to overrate it, but it is a hint that different forms of collaboration may be reflected differently on the World Wide Web. We also investigated the influences of different sub-disciplines and team maturity on the inlinks to the web pages. However, as the results were not significant or showed no clear pattern we omitted them.

Summarising the bivariate analyses of the site inlinks (inlinks from other domains) to the web pages of life sciences research teams we conclude, that size – of the team and of the web presence – is related to the site inlinks and therefore the visibility of a team on the web. In addition, we find also a positive relationship between research productivity and collaborations, in particular international and business collaborations, and site inlinks. This applies only to the staff-normalised link counts that are recommended in the literature, however, and none of the analyses with page-normalised link counts is significant.

Results of the multivariate analyses

In order to take into account the interaction between different factors which might explain differences in the numbers of site inlinks that a web site receives, we computed multivariate regressions. These regressions use the link counts as the dependent variables and the page counts, as well as certain characteristics of the teams and their team leaders as the independent variables. Several combinations of these variables were estimated. The following results show the best performing models with the most stable results (see above for a more detailed description of the estimation methods).

Table 8 reproduces the results for the raw link counts with all three search engines. The following results appear remarkable from our point of view:

- Sweden was chosen as the reference country in the estimations, because its link numbers were closest to the overall ten-country average in Table 4. Thus we expected that some countries have higher and others have lower coefficients in the estimations. However, in the estimated models for the inlinks retrieved with Google nearly all countries have negative coefficients, i.e. lower link scores than Sweden (model 1 in Table 8); and for five out of nine countries – Czech Republic, Spain, Italy, Norway, and Portugal – this is

statistically significant. No country has a positive and significant coefficient (= significantly more inlinks than to Swedish teams). The Swedish teams perform well in the Yahoo! data (model 3), too, with six out of nine countries having significantly lower coefficients (the same countries as before plus France). But in the Yahoo! web data three countries, Germany, Hungary and the UK, have significantly higher coefficients than Sweden. For both, Google and Yahoo! data most coefficients are reduced and become insignificant if further variables are included (models 2 and 4). Using the MSN link scores in the estimations as the dependent variable, only Portugal has significantly fewer site inlinks to its life science teams than Sweden, and the UK, Germany and Hungary still have significantly more inlinks (model 5 in Table 8). These differences point to differences in the search engine coverage between the countries. They indicate that Google has better coverage of pages linking to Swedish life sciences websites and Yahoo! and MSN have better coverage of pages linking to British and Hungarian sites. This is also confirmed in Table 4 (above) which shows that teams from Sweden receive above average inlinks according to Google and below average inlinks according to MSN.

- The number of pages is clearly the most important predictor of the site inlinks. We used the logged link page count with each search engine to avoid estimation problems which resulted from the size of some sites, e.g. the maximum page count for Google was 418,000 and for Yahoo! was 81,300. This somewhat distorts the relationship of this coefficient to the other coefficients, but without doubt the number of pages is the most important and significant cause of site inlinks for all search engines.
- In addition, a very small but also significant effect is discernible for the team size: the larger a team, the more inlinks it received to its web pages, even if the number of pages was held constant. In models 1, 2 and 5 this effect is captured partially by the publication variable that is correlated to team size, too.
- The publication output of a team was insignificant in all estimations except for model 1. The variable for the mean number of citations per publication (TOTMORC), an indicator for the quality of the publications was also mostly insignificant. Only in the MSN estimations (model 6) it nearly reaches the 5%-significance level. However, the coefficient is not positive as we would expect assuming that authors of higher quality publications also receive more external inlinks to their web presences, but it is negative.
- The extent of research collaboration with other teams was assessed in several different ways – in total, at international level, with private enterprises – which mostly did not lead

to any measurable effect on site inlinks. The coefficients for university-industry partnerships leading to a funding of PhD students or post-docs are also negative and not as expected. They are only significant for the Google link counts, but still the negative tendency seems clear.

- The coefficients for team age, the level of recognition of the team leaders and gender of the team leader are insignificant in all estimations. We also included dummy variables for life sciences sub-disciplines which were never significant and therefore omitted from the estimations.

Table 8: Estimation results for external link counts with Google (GO), Yahoo! (YA) and MSN (MS)

[Table 8 about here]

In addition to these regressions of the raw link count from each search engine, we computed further regressions with the logged maximum link count: for each research team the maximum number of site inlinks from other domains retrieved by one of the three search engines Google, Yahoo!, and MSN was taken as the dependent variable (Table 9). The OLS regressions were subject to heteroscedasticity, as the Breusch-Pagan-Test shows. Hence, we computed models employing two different types of corrections, the White estimator (models 1 & 2) and Weighted Least Squares (models 3 & 4). Moreover, we estimated Tobit models to reduce the influences of the non-normality of the residuals (models 5 & 6). The results of the estimations are quite similar and similar to the more extensively discussed results from the raw link count models with few exceptions. The weaknesses of country coverage for individual search engines seem to have been eliminated. We can distinguish two groups of countries: Czech Republic, France, Spain, Italy and Portugal where teams had rather few site inlinks, compared to the UK, Hungary, Germany, Norway and Sweden where teams had more site inlinks. The size of the web presence and the size of the research group are corroborated as significant predictors for site inlinks. Moreover, the negative coefficients for the number of citations per publication and industry funding of PhD students or post-docs are confirmed.

Table 9: Estimation results for logged maximum external link counts

[Table 9 about here]

Discussion and conclusions

The analysis produced some interesting insights into the performance of different search engines, the consequences of factoring out size effects in different ways, and the visibility of life science research teams from different countries on the web:

- a) Google retrieved fewer site inlinks to research teams in the UK and Hungary and MSN retrieved fewer links to Swedish teams. The national differences in inlink counts found in some of the statistical analyses were unexpected and hence pursued to identify a cause. Although national differences in search engine coverage of the web are known for individual search engines, the underlying causes seemed to be due to the age of the sites in different countries, suggesting that the bias should be the same for each country (Vaughan, & Thelwall, 2004). Our data included significant variations between search engines, for example whilst MSN's inlink counts for Sweden were, on average, 5.4 times as big as Google's, its counts for the UK were 18 times as big. There are three possible explanations for inter-search engine variations. First, Google counts only links to the home page and hence it is possible that UK sites tended to have more links to pages other than the home page, for example because they hosted significant resources. This appeared to be true in our data to some extent, but not sufficiently to explain the Google/MSN discrepancy. Second, links to UK sites tend to come disproportionately from two of the oldest web domains, .edu and .ac.uk, but this should favour the older search engine Google, which had the extra time to index .ac.uk and .edu sites. Third, one of the search engines may use a non-linear algorithm and report disproportionately many/few links for larger sites. We have no evidence for this, but it would explain the results. Logically, since Google only reports a fraction of the links it knows about and MSN appears to report all or most of its known links, this factor is most likely to affect Google. Hence we suggest, albeit without proof, that Google's link counts are unreliable for statistical analyses using assumptions of linearity.

The problem could be reduced, if not solved, through taking the maximum number of site inlinks retrieved from any of the three search engines. This does not resolve all doubt whether teams in countries with few inlinks actually receive more inlinks on pages not retrieved by the search engines. However, site inlinks from pages that are not retrieved by one of the major search engines are probably not really increasing the visibility of a team on the web either.

- b) If the influence of size is factored out in order to obtain a size-neutral link count, it matters a lot which denominator is chosen. Using the page count (an indicator of the size of the

web presence) changes the picture entirely whereas using the number of scientific staff (a team size indicator) produces a similar country ranking as the unstandardised data. However, further information on the denominator used for normalisation is also needed: Do the web publishing strategies and the amount of content included on a web page differ across countries?

- c) Assuming that most systematic differences for retrieved inlinks that are due to variations in search engine coverage were compensated through using data from three major search engines, we can separate the countries in roughly two groups: the UK, Hungary, Germany, Norway and Sweden where teams had more site inlinks and the Czech Republic, France, Spain, Italy and Portugal where teams had fewer site inlinks. The web visibility is higher for teams from the former countries and lower for teams from the latter countries.

Team size had a significant positive but very small effect on the site inlinks. Larger teams produce more pages and they are more visible on the web. But an additional scientist on average leads to only 0.005 additional site inlinks which is negligible from a quantitative perspective. Similarly, an additional page also causes less than one additional link, but – using again the estimated coefficients from Table 9 – in this case it is roughly 0.4 additional links per additional page, or 4 links for 10 pages. These calculations stress again that both the size of the web presence and the size of the team are important for explaining the number of site inlinks, but web content size is clearly more influential.

In regard to the included output, productivity and research quality indicators we got results which are only partially in line with previous findings. As in a previous paper with a different dataset (Barjak, Li, & Thelwall, in press) we only obtained small correlation coefficients between research output and site inlinks and we did not find any significant coefficients in multivariate analyses which factored out the size effects. The coefficient for the quality of publications (measured through the mean number of citations per publication) is not as expected and negative. We expected to get a positive coefficient, as papers of high quality should raise the community's interest in the web sites of the authors and due to previous findings for academic departments (Li, Thelwall, Wilkinson, & Musgrove, 2005; Tang & Thelwall, 2003). A possible cause could be that highly cited papers are posted less often on the web, because they appear more often in prestigious journals which restrict the self-publishing of articles on the web and because teams do not have to republish them to gain additional visibility. However, this is not in line with previous findings in computer science which point to more citations for articles available online (Lawrence, 2001). We do not have a

satisfactory explanation for this result and the relationship certainly needs to be investigated further.

The collaboration indicators show some weak relationships in the bivariate analyses which are insignificant in the multivariate models. The rather anomalous negative relationship to site inlinks obtained for the homepages of scientists from five different disciplines and seven European countries (Barjak, Li, & Thelwall, in press) is thus not corroborated. But why don't we get a positive correlation? We included the teams' entire web sites in the link retrieval with Yahoo! and MSN. Links to project pages which may be more likely to be hyperlinked by collaborators should be included. Again we do not have a satisfactory explanation. Of course, the collaboration indicators cover only a very small sector of collaborations related mostly to research. Collaborations with sponsors, governmental offices or professional associations, as well as collaborations referring to other tasks such as teaching, consulting and technology transfer, or administration are not included. Hence, we might need better collaboration indicators. The only collaboration indicator that is not directly related to research, but based on whether PhD students or post-docs were funded by money from industry, produced a negative coefficient. Teams which had this type of relationship to industry received fewer external inlinks.

Characteristics of the team leader: We included the gender and the level of professional recognition of the team leader as additional variables as we have shown previously that they might influence the number of site inlinks (the age of the team leader was not included as we used the age of the team instead) (Barjak, Li, & Thelwall, in press). However, we do not obtain significant effects either for gender or for the level of professional recognition. Though the gender results all point to more links for male-led teams, this result is never significant in the multivariate estimations and probably a reflection of size – both, size of the web presence and team size – in the bivariate calculations. We obtained a similar result for the level of professional recognition. Well recognised scientists and male scientists have larger teams and produce more web content leading to more site inlinks, but they do not receive more inlinks because of their high recognition or gender.

Acknowledgement

This study uses data obtained within the NetReAct project ("The role of Networking in Research Activities") commissioned by the Institute for Prospective Technological Studies of the European Commission's Joint Research Centre.

References

- Adamic, L. A., & Huberman, B. A. (2000). Power-law distribution of the World Wide Web. *Science*, 287(24 March), 2115a.
- Allen, T. (1991). *Managing the flow of technology: technology transfer and the dissemination of technological information within the R&D organization* (5 ed.). Cambridge MA: MIT-Press.
- Barabási, A. L., & Albert, R. (1999). Emergence of scaling in random networks. *Science*, 286, 509-512.
- Bar-Ilan, J. (1999). *Search engine results over time - a case study on search engine stability*. Retrieved January 26, 2006, from <http://www.cindoc.csic.es/cybermetrics/articles/v2i1p1.html>
- Bar-Ilan, J. (2004). The use of Web search engines in information science research. *Annual Review of Information Science and Technology*, 38, 231-288.
- Bar-Ilan, J. (2004). A microscopic link analysis of academic institutions within a country - the case of Israel. *Scientometrics*, 59(3), 391-403.
- Bar-Ilan, J. (2005). What do we know about links and linking? A framework for studying links in academic environments. *Information Processing & Management*, 41(4), 973-986.
- Barjak, F. (2006). Research productivity in the internet era. *Scientometrics*, 68(3), 343-360.
- Barjak, F., Li, X., & Thelwall, M. (in press). Which factors explain the web impact of scientists' personal homepages? *Journal of the American Society for Information Science and Technology*.
- Beaulieu, A., & Simakova, E. (2006). Textured Connectivity: an ethnographic approach to understanding the timescape of hyperlinks [Electronic Version]. *Cybermetrics*, 10 from <http://www.cindoc.csic.es/cybermetrics/articles/v10i1p6.html>.

- Broder, A., Kumar, R., Maghoul, F., Raghavan, P., Rajagopalan, S., Stata, R., et al. (2000). Graph structure in the web. *Journal of Computer Networks*, 33(1-6), 309-320.
- Cameron, A. C., & Trivedi, P. K. (1998). *Regression analysis of count data*. Cambridge: Cambridge University Press.
- Chen, C., Newman, J., Newman, R., & Rada, R. (1998). How did university departments interweave the Web: A study of connectivity and underlying factors. *Interacting with Computers*, 10, 353-373.
- Cummings, J. N., & Kiesler, S. (2003). KDI initiative: Multidisciplinary scientific collaboration. Retrieved 17 May, 2005, from http://netvis.mit.edu/papers/NSF_KDI_report.pdf
- Etzkowitz, H., Kemelgor, C., Neuschatz, M., & Uzzi, B. (1992). Athena unbound: Barriers to women in academic science and engineering. *Science and Public Policy*, 19(3), 157-179.
- Fogg, B. J., Marshall, J., Laraki, O., Osipovich, A., Varma, C., Fang, N., et al. (2001). *What makes Web sites credible?: a report on a large quantitative study*. Paper presented at the SIGCHI conference on Human Factors in Computing Systems, Seattle, Washington.
- Fry, J. (2006). Studying the Scholarly web: How disciplinary culture shapes online representations [Electronic Version]. *Cybermetrics*, 10. Retrieved 21 April 2006 from <http://www.cindoc.csic.es/cybermetrics/articles/v10i1p2.html>.
- Greene, W. H. (2000). *Econometric Analysis* (4 ed.). Upper Saddle River, NJ: Prentice Hall.
- Harries, G., Wilkinson, D., Price, L., Fairclough, R., & Thelwall, M. (2004). Hyperlinks as a data source for science mapping. *Journal of Information Science*, 30(5), 436-447.
- Heimeriks, G., & Besselaar, P. V. d. (2006). Analyzing hyperlinks networks: The meaning of hyperlink based indicators of knowledge [Electronic Version]. *Cybermetrics*, 10. Retrieved 21 April 2006 from <http://www.cindoc.csic.es/cybermetrics/articles/v10i1p1.html>.
- 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. (2004). *New media and communication networks in knowledge production: a case study*.

- Ingwersen, P. (1998). The Calculation of Web Impact Factors. *Journal of Documentation*, 54(2), 236-243.
- Kleinberg, J. (1999). Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 46(5), 604-632.
- Kraut, R., Egido, C., & Galegher, J. (1990). Patterns of contact and communication in scientific research collaboration. In R. Kraut, C. Egido & J. Galegher (Eds.), *Intellectual teamwork Social and technological foundations of cooperative work* (1 ed., pp. 149-171). Hillsdale: Lawrence Erlbaum Associates.
- Lawrence, S. (2001). Free online availability substantially increases a paper's impact. *Nature*, 411(6837), 521.
- Lawrence, S., & Giles, C. L. (1999). Accessibility of information on the web. *Nature*, 400, 107-109.
- Li, X., Thelwall, M., Musgrove, P., & Wilkinson, D. (2003). The relationship between the links/Web Impact Factors of computer science departments in UK and their RAE (Research Assessment Exercise) ranking in 2001. *Scientometrics*, 57(2), 239-255.
- Li, X., Thelwall, M., Wilkinson, D., & Musgrove, P. (2005). 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. (2005). National and international university departmental Web site interlinking. Part 2: Link patterns. *Scientometrics*, 64(2), 187-208.
- Liu, Z. (2004). Perceptions of credibility of scholarly information on the web. *Information Processing & Management*, 40(6), 1027-1038.
- 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
- Mettrop, W., & Nieuwenhuysen, P. (2001). Internet search engines - fluctuations in document accessibility. *Journal of Documentation*, 57(5), 623-651.
- Nelson, M. (2005). Academic home pages and Nobel laureates. In P. Ingwersen & B. Larsen (Eds.), *Proceedings of ISSI 2005 - 10th International Conference of the International*

Society for Scientometrics and Informetrics (Vol. 1, pp. 193-196). Stockholm, Sweden: Karolinska University Press.

Park, H. W., Barnett, G. A., & Nam, I.-Y. (2002). Hyperlink-affiliation network structure of top web sites: examining affiliates with hyperlink in Korea. *Journal of the American Society for Information Science and Technology*, 53(7), 592-601.

Pennock, D., 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. *Proceedings of the National Academy of Sciences*, 99, 5207-5211.

Prpic, K. (2002). Gender and productivity differentials in science. *Scientometrics*, 55(1), 27-58.

Rousseau, R. (1992). Why am I not cited or why are multi-authored papers more cited than others? *Journal of Documentation*, 48(1), 79-80.

Rousseau, R. (1999). Daily time series of common single word searches in AltaVista and NorthernLight. *Cybermetrics*, 2/3, Retrieved July 25, 2006 from: <http://www.cindoc.csic.es/cybermetrics/articles/v2002i2001p2002.html>.

Shauman, K., & Xie, Y. (2003). Explaining Sex Differences in Publication Productivity among Postsecondary Faculty. In L. S. Hornig (Ed.), *Equal Rights, Unequal Outcomes Women in American Research Universities* (1 ed., pp. 175-208). New York: Kluwer Academic/Plenum Publishers.

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

Smith, A., & Thelwall, M. (2005). Web links as an indicator of research output: a comparison of NZ Tertiary Institution links with the Performance Based Research Funding assessment. In P. Ingwersen & B. Larsen (Eds.), *Proceedings of ISSI 2005 - the 10th International Conference of the International Society for Scientometrics and Informetrics* (Vol. 1, pp. 205-211). Stockholm: Karolinska University Press.

Snyder, H. W., & 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). Disciplinary differences in US academic departmental Web site interlinking. *Library and Information Science Research*, 25(4), 437-458.

- Tang, R., & Thelwall, M. (2004). Patterns of national and international web inlinks to US academic departments: An analysis of disciplinary variations. *Scientometrics*, 60(3), 475-485.
- Thelwall, M. (2001). Extracting Macroscopic Information from Web Links. *Journal of the American Society for Information Science and Technology*, 52(13), 1157-1168.
- 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. (2002c). An initial exploration of the link relationship between UK university websites. *ASLIB Proceedings*, 54(2), 118-126.
- Thelwall, M. (2002d). Methodologies for crawler based web surveys. *Internet Research: Electronic Networking and Applications*, 12(2), 124-138.
- Thelwall, M. (2003). What is this link doing here? Beginning a fine-grained process of identifying reasons for academic hyperlink creation. *Information research*, 8(3), no. 151.
- Thelwall, M. (2004). *Link analysis: an information science approach*. Amsterdam: Elsevier Academic Press.
- Thelwall, M. (2006). Interpreting Social Science Link Analysis Research: A Theoretical Framework. *Journal of the American Society for Information Science and Technology*, 57(1), 60-68.
- Thelwall, M., Barjak, F., & Kretschmer, H. (2006). Web links and gender in science: An exploratory analysis. *Scientometrics*, 67(3), 373-383.
- Thelwall, M., & Harries, G. (2003). The Connection between the Research of a University and Counts of Links to its Web Pages: An Investigation Based Upon a Classification of the Relationships of Pages to the Research of the Host University. *Journal of the American Society for Information Science and Technology*, 54(7), 594-602.
- 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., Harries, G., & Wilkinson, D. (2003). Why do web sites from different academic subjects interlink? *Journal of Information Science*, 29(6), 453-471.
- Thelwall, M., Vaughan, L., Cothey, V., Li, X., & Smith, A. G. (2003). Which academic subjects have most online impact? A pilot study and a new classification process. *Online Information Review*, 27(5), 333-343.
- Thelwall, M., & Wilkinson, D. (2003). Three Target Document Range Metrics for University Web Sites. *Journal of the American Society for Information Science and Technology*, 54(6), 489-496.
- Vaughan, L., & Thelwall, M. (2004). Search engine coverage bias: evidence and possible causes. *Information Processing & Management*, 40(4), 693-707.
- 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.

Appendix

Table: Variables of the estimations

Acronym	Variable	Mean	S. Dev.
Dependent Variables			
EXT_GOO	Google site inlinks to home page	2.0	
EXT_YAH	Yahoo! Site inlinks	33.7	
EXT_MSN	MSN Site Inlinks	15.2	
LEXT_MAX	Logged maximum site inlinks		
Independent Variables			
PAGE_GOO	Google page count	1008.0	
PAGE_YAH	Yahoo! page count	250.7	
PAGE_MSN	MSN page count	40.3	
LPAG_MAX	Logged maximum page count		
F9_SCI	Team size (total scientific staff)		
CZ, DE, ES, FR, HU, IT, NO, PT, SE, UK	Country dummies: Czech Republic, Germany, Spain, France, Hungary, Italy, Norway, Portugal, Sweden, United Kingdom	–	–
TEAMAGE	Team age	9.803	8.526
AVRECOG	Average recognition of the team head (dummy)	–	–
HIRECOG	High recognition of the team head (dummy)	–	–
GENDER	Gender of the team head (Male = 0, Female = 1)	–	–
TOTPAP/ ZTOTPAP	Total number of papers published in the Thomson ISI SCIE database in 2001 (ZTOTPAP: per team member)	5.350	5.833
TOTMOCR	Mean Observed Citation Rate 2001-2003 of the 2001 papers	6.253	7.076
COLLABTO	Total number of collaborating research teams per team member		
ICPAP/ZICPAP	Papers with foreign co-authors published in the Thomson ISI SCIE database in 2001 (ZICPAP: per team member)	0.147	0.268
BCPAP01	Papers with co-authors from industry published in the Thomson ISI SCIE database in 2001 (dummy)	–	–
INDUFUND	Industrial funding of PhD students or post-docs (dummy)	–	–

Figure 1

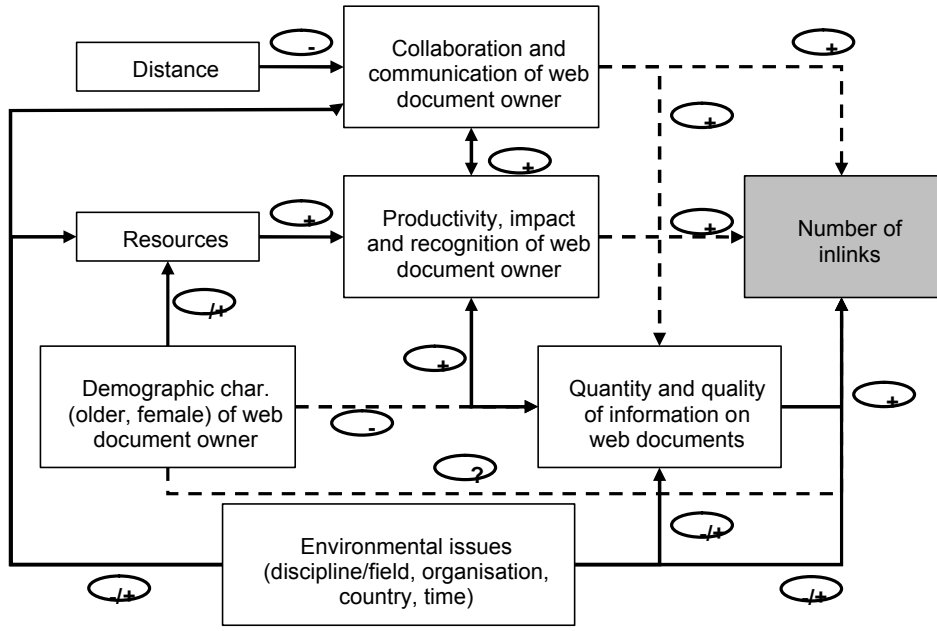


Table 1:

Engine	Page counts	All inlinks	Site inlinks
Google (site with its own domain name)	site:d	link:http://d	link:http://d (& manual filtering)
Google (site without its own domain name)	inurl:pg site:d	link:http://u	link:http://u (& manual filtering)
MSN (site with its own domain name)	site:d	linkdomain:d	linkdomain:d -site:d
MSN (site without its own domain name)	inurl:pm site:d (& manual rechecking if zero matches)	link:http://u	link:http://u -site:d
Yahoo! (site with its own domain name)	site:d	linkdomain:d	linkdomain:d -site:d
Yahoo! (site without its own domain name)	inurl:py site:d (& manual checking of matches)	link:http://u	link:http://u -site:d

Table 2 :

	Mean	Median
Google page count	1008.0	1
Yahoo! page count	250.7	2
MSN page count	40.3	1
Google home page inlinks	6.8	1
Yahoo! all inlinks	134.3	9
MSN all inlinks	35.4	2
Google site inlinks to home page	2.0	0
Yahoo! site inlinks	33.7	1
MSN site Inlinks	15.2	1

Table 3

	Google Page Count	Yahoo! Page Count	MSN Page Counts	Google home page inlinks	Yahoo! All Inlinks	MSN All Inlinks	Google site inlinks to home page	Yahoo! Site inlinks	MSN Site Inlinks
Google Page Count	1	0.86	0.77	0.42	0.54	0.43	0.41	0.52	0.45
Yahoo! Page Count		1	0.75	0.44	0.57	0.46	0.43	0.57	0.5
MSN Page Counts			1	0.46	0.56	0.48	0.41	0.51	0.47
Google home page inlinks				1	0.75	0.61	0.6	0.54	0.56
Yahoo! All Inlinks					1	0.74	0.52	0.75	0.67
MSN All Inlinks						1	0.63	0.79	0.9
Google site inlinks to home page							1	0.67	0.65
Yahoo! Site inlinks								1	0.84
MSN Site Inlinks									1

Table 4

Country	Google site inlinks to home page	Yahoo! Site inlinks	MSN Site Inlinks	Logged maximum page count ^a	Logged maximum site inlinks ^a	Page-normalised link count ^b	Staff-normalised link count ^c
CZ	0.7 (0.4)	14.3 (6.0)	5.5 (2.7)	1.0 (0.2)	0.5 (0.1)	-0.5 (0.1)	-0.5 (0.1)
DE	3.2 (0.8)	28.7 (8.8)	8.3 (2.9)	1.5 (0.1)	0.9 (0.1)	-0.6 (0.1)	0.0 (0.1)
ES	0.3 (0.1)	4.2 (1.8)	2.2 (0.9)	0.7 (0.1)	0.3 (0.1)	-0.4 (0.1)	-0.6 (0.1)
FR	1.9 (0.7)	8.7 (2.7)	6.2 (2.3)	0.9 (0.1)	0.5 (0.1)	-0.3 (0.1)	-0.5 (0.1)
HU	3.0 (1.4)	62.6 (28.8)	20.4 (14.5)	1.1 (0.2)	0.8 (0.2)	-0.4 (0.1)	-0.1 (0.1)
IT	0.3 (0.1)	2.5 (1.0)	1.7 (0.7)	0.6 (0.1)	0.3 (0.1)	-0.3 (0.1)	-0.6 (0.1)
NO	1.2 (0.8)	10.8 (9.2)	3.1 (1.7)	0.7 (0.1)	0.4 (0.1)	-0.2 (0.1)	-0.4 (0.1)
PT	1.0 (0.3)	9.8 (2.9)	4.3 (1.4)	1.0 (0.1)	0.6 (0.1)	-0.5 (0.1)	-0.5 (0.1)
SE	2.9 (1.2)	20.7 (10.9)	10.2 (5.9)	0.9 (0.1)	0.6 (0.1)	-0.3 (0.1)	-0.3 (0.1)
UK	3.1 (1.1)	112.4 (59.9)	56.7 (29.1)	1.0 (0.1)	0.9 (0.1)	-0.1 (0.1)	-0.1 (0.1)
Total	2.0 (0.3)	33.7 (10.7)	15.2 (5.2)	1.0 (0.0)	0.6 (0.0)	-0.3 (0.0)	-0.3 (0.0)
F-stat./Chi ^d	1.9+/38.2**	1.5/47.9**	1.6/40.6**	5.0**/41.8**	5.7**/48.2**	2.4*/26.9**	5.6**/48.2**

Table 5

	Google site inlinks to home page	Yahoo! Site inlinks	MSN Site Inlinks	Log page count (max.)	Log site inlinks (max.)	Page-normalised link count ^a	Staff-normalised link count ^b
Total publications	0.07/0.15**	-0.02/0.13*	0.02/0.18**	0.15**/0.12*	0.13*/0.15**	-0.07/-0.02	0.01/0.01
Publications per scientist	0.02/0.03	-0.03/0.00	-0.03/0.06	0.04/0.02	0.01/0.03	-0.04/-0.01	0.10*/0.14**
Total citations	0.03/0.12*	-0.02/0.07	0.05/0.08	0.10+/-0.13*	0.03/0.08	-0.09+/-0.08	0.00/0.01
Citations per scientist	-0.03/0.00	-0.05/-0.05	-0.03/-0.05	0.00/0.03	-0.08/-0.04	-0.08/-0.06	0.02/0.09
Citations per publication	-0.01/0.06	-0.04/0.03	0.00/-0.01	-0.01/0.08	-0.05/0.02	-0.03/-0.05	-0.04/0.02

Table 6

	Google site inlinks to home page	Yahoo! Site inlinks	MSN Site Inlinks	Log page count (max.)	Log site inlinks (max.)	Page-normalised link count ^a	Staff-normalised link count ^b
Number of collaborating teams	0.09+/0.08+	0.00/0.09*	0.07+/0.11*	0.05/0.07	0.09+/0.09+	0.03/-0.01	-0.05/-0.06
Collaborating teams per scientist	-0.04/-0.06	-0.03/-0.10*	-0.01/-0.07	-0.10*/-0.11*	-0.05/-0.09+	0.07/0.04	0.11*/0.14*
Publications with co-authors from other teams	0.04/0.06	-0.03/0.04	0.01/0.09+	0.13*/0.09+	0.11*/0.07	-0.06/-0.06	0.01/-0.04
Publications with co-authors from other teams per scientist	-0.01/-0.05	-0.03/-0.09+	-0.03/-0.05	-0.01/-0.02	-0.05/-0.06	-0.04/-0.04	0.08/0.06
Publications with co-authors from foreign teams	0.05/0.08+	0.00/0.11*	0.04/0.14**	0.15**/0.12*	0.13*/0.13**	-0.07/-0.03	0.03/0.02
Publications with co-authors from foreign teams per scientist	0.02/0.02	0.00/0.05	-0.01/0.08	0.05/0.06	0.02/0.07	-0.04/-0.01	0.10*/0.08+

Table 7

	Google site inlinks to home page	Yahoo! Site inlinks	MSN Site Inlinks	Log page count (max.)	Log site inlinks (max.)	Page-normalised link count ^b	Staff-normalised link count ^c
Male team leader	2.1 (0.3)	38.0 (13.2)	17.5 (6.4)	1.0 (0.0)	0.7 (0.0)	-0.3 (0.0)	-0.3 (0.0)
Female team leader	1.2 (0.4)	15.8 (5.6)	5.9 (2.8)	0.9 (0.1)	0.5 (0.1)	-0.3 (0.1)	-0.3 (0.1)
F-stat./Chi ^d	1.7/6.5*	0.7/2.2	0.8/5.0*	1.0/0.3	1.8/2.0	0.0/0.4	0.1/0.0

Table 8

	Model 1 (GO)		Model 2 (GO)		Model 3 (YA)		Model 4 (YA)		Model 5
	Negbin, ZIP=normal		Negbin, ZIP=normal		Negbin, ZIP=normal		Negbin, ZIP=normal		Negbin, ZIP
	Coeff.	t-ratio	Coeff.	t-ratio	Coeff.	t-ratio	Coeff.	t-ratio	Coeff.
Constant	-0.01	-0.11	-0.22	-0.71	0.64	2.51*	0.13	0.28	-0.07
CZ	-1.23	-4.15**	-1.20	-1.41	-0.55	-1.98*	-0.17	-0.29	0.01
DE	0.26	1.19	0.53	1.94+	0.58	2.73**	0.11	0.29	0.49
ES	-0.60	-3.90**	-0.28	-1.20	-0.99	-3.47**	-0.40	-0.54	-0.05
FR	-0.15	-1.12	0.65	1.59	-0.83	-3.00**	-0.56	-1.38	-0.01
HU	-0.28	-1.82+	-0.23	-1.04	0.57	1.92+	1.21	3.03**	0.61
IT	-0.38	-2.70**	-0.25	-0.86	-1.15	-4.56**	-0.95	-2.73**	-0.44
PT	-0.80	-3.88**	-0.47	-1.69+	-0.75	-2.44*	-0.42	-0.86	-0.52
NO	-0.27	-1.90+	-0.04	-0.13	-0.88	-3.07**	-0.69	-1.12	-0.08
UK	-0.01	-0.06	0.08	0.21	0.60	2.43*	1.01	2.85**	0.97
LPAGE	0.28	7.43**	0.25	5.05**	0.66	11.11**	0.71	9.55**	0.66
F9_SCI	2.52E-03	0.78	3.58E-03	0.23	0.03	3.07**	0.04	2.75**	0.01
TOTPAPNE	0.03	3.14**	2.27E-02	0.84	-0.01	-0.47	-0.02	-0.73	0.02
TOTMOCR			-4.09E-03	-0.47			-0.01	-0.33	
COLLABTO			-2.32E-04	-0.01			-0.01	-0.47	
TEAMAGE			1.13E-02	1.22			0.01	0.55	
AVRECOG			-4.41E-04	0.00			0.22	0.57	
HIRECOG			-0.09	-0.40			-0.10	-0.24	
GENDER			-0.10	-0.72			-0.03	-0.11	
ICPAP01			0.05	0.26			0.13	0.28	
INDUFUND			-0.49	-1.60			-0.17	-0.29	
BCPAP01			-0.06	-0.22			0.11	0.29	
Alpha	2.89	13.86**	3.24	7.10**	2.96	38.56**	2.68	23.95**	2.46
Log-L	-558.50		-443.70		-1108.38		-881.97		-892.60
Log-L(0)	-561.48		-454.61		-1121.23		-922.62		-906.37
VuongSt	0.69		1.69		2.59		5.07		3.62
Tau	-3.51	-3.75**	-347.63	-0.09	-1.21	-3.69**	-1715.19	-0.01	-7124.34
Cases	408		323		408		323		408

significance levels ** = 0.01, * = 0.05, + = 0.1.

Table 9

	OLS, White		OLS, White		WLS (lpag_max)		WLS (lpag_max)		TOBI
	Parameter	t-ratio	Parameter	t-ratio	Parameter	t-ratio	Parameter	t-ratio	Parame
Constant	0.14	1.69+	0.22	1.60	0.08	0.62	0.13	0.65	-0.13
CZ	-0.19	-1.60	-0.26	-1.62	-0.31	-1.98*	-0.41	-2.14*	-0.30
DE	-0.01	-0.10	-0.07	-0.50	-0.19	-1.51	-0.25	-1.65+	0.02
ES	-0.25	-2.51*	-0.21	-1.89+	-0.33	-2.02*	-0.28	-1.46	-0.41
FR	-0.14	-1.37	-0.14	-1.29	-0.22	-1.60	-0.28	-1.62	-0.20
HU	0.00	0.00	0.07	0.42	-0.01	-0.09	0.03	0.14	-0.05
IT	-0.22	-2.40*	-0.26	-2.36*	-0.27	-1.67+	-0.32	-1.58	-0.35
PT	-0.18	-1.65	-0.26	-2.04*	-0.18	-1.26	-0.31	-1.71+	-0.32
NO	-0.09	-0.87	-0.08	-0.64	-0.11	-0.64	-0.15	-0.77	-0.06
UK	0.18	1.66+	0.19	1.47	0.15	1.14	0.10	0.63	0.26
LPAG_MAX	0.47	12.75**	0.47	11.44**	0.53	16.10**	0.51	12.90**	0.58
F9_SCI	0.01	2.70**	0.01	2.43*	0.01	3.49**	0.01	2.65**	0.01
ZTOTPAP	0.03	0.62	-0.01	-0.16	0.01	0.26	-0.03	-0.42	0.05
TOTMOCR			-6.19E-03	-3.20**			-8.43E-03	-2.14*	
COLLABTO			3.34E-04	0.09			8.59E-03	1.57	
TEAMAGE			7.38E-04	0.20			4.31E-03	0.95	
AVRECOG			-0.05	-0.70			-0.05	-0.49	
HIRECOG			-0.04	-0.49			-0.03	-0.31	
GENDER			-0.01	-0.13			-0.08	-0.82	
ICPAP01			0.06	0.72			0.06	0.66	
INDUFUND			-0.20	-2.26*			-0.35	-3.12**	
BCPAP01			0.11	1.28			0.13	1.25	
Adj. Rsquared	0.45		0.46		0.45		0.49		-
F	29.14**		12.93**		29.32**		14.40**		-
Normality test	BS chi	19.93**	BS chi	7.67*	BS chi	14.24**	BS chi	4.19	CM of
Het. test	BP	83.51**	BP	71.64**	-		-		LR
Cases	443		341		443		341		443

significance levels ** = 0.01, * = 0.05, + = 0.1.