



ELSEVIER

Available online at www.sciencedirect.com

SCIENCE @ DIRECT®

Information Processing and Management 41 (2005) 1495–1510

**INFORMATION
PROCESSING
&
MANAGEMENT**

www.elsevier.com/locate/infoproman

Mathematical models for academic webs: Linear relationship or non-linear power law?

Nigel Payne *, Mike Thelwall

School of Computing and Information Technology, University of Wolverhampton, 35/49 Lichfield Street, Wolverhampton WV1 1EQ, UK

Received 27 September 2004; accepted 3 March 2005
Available online 20 April 2005

Abstract

Previous studies of academic web interlinking have tended to hypothesise that the relationship between the research of a university and links to or from its web site should follow a linear trend, yet the typical distribution of web data, in general, seems to be a non-linear power law. This paper assesses whether a linear trend or a power law is the most appropriate method with which to model the relationship between research and web site size or outlinks. Following linear regression, analysis of the confidence intervals for the logarithmic graphs, and analysis of the outliers, the results suggest that a linear trend is more appropriate than a non-linear power law.

© 2005 Elsevier Ltd. All rights reserved.

Keywords: Web links; Academic webs; Linking models; Linear relation; Non-linear power law

1. Introduction

Research involving academic web links is a relatively new area of study, part of webometrics (Almind & Ingwersen, 1997; Björneborn & Ingwersen, 2004), but is one that has already produced many substantial contributions, primarily through methodological developments and the uncovering of mathematical patterns and relationships hidden within the structure of the hyperlinks.

Hyperlinking is a fundamental element of web use, and links between sites with substantial academic content, in particular, are often viewed as analogues of citations in scientific journals (Bar-Ilan, 2004; Cronin, 2001; Larson, 1996; Rousseau, 1997). Nevertheless, analogies have to be treated with caution, both in

* Corresponding author.

E-mail addresses: n.c.payne@wlv.ac.uk (N. Payne), m.thelwall@wlv.ac.uk (M. Thelwall).

terms of data collection and the interpretation of results (Bar-Ilan, 2001; Björneborn & Ingwersen, 2001; Egghe, 2000; Prime, Bassecouard, & Zitt, 2002; van Raan, 2001). In Björneborn's webometrics terminology (Björneborn & Ingwersen, 2004), links to a web site are (*site*) *inlinks*, links from a site are (*site*) *outlinks*, and links between pages in the same site are (*site*) *self-links*. The term *interlinking* will be used for links between a pair of web sites or within a set of web sites, excluding site self-links.

Several studies have investigated the relationship between interlinking counts of a selected set of universities on a national level and the research productivity of those universities. Positive correlations were found between research measures and inlink counts for universities in the United Kingdom (Thelwall, 2001a, 2002a), Australia (Smith & Thelwall, 2002), Taiwan and Mainland China (Thelwall & Tang, 2003), and Canada (Vaughan & Thelwall, 2005). More specifically, counts of links to universities seem to be approximately proportional to the quadruple product of the size in academic staff numbers and research quality of the source and target institutions (Thelwall, 2002d). This is consistent with the finding that universities with higher research productivity attract more links because they create more web pages, rather than because their individual pages are more likely to attract links (Thelwall & Harries, 2004). It is important to bear in mind that although link counts correlate strongly with research, university web site link targets (from HTML pages) are very rarely academic papers (Wilkinson, Harries, Thelwall, & Price, 2003), although many more links to academic papers are present in postscript and PDF documents on the web (Goodrum, McCain, Lawrence, & Giles, 2001). Approximately 90% of links between national universities are created for reasons associated with educational and scholarly activity (Wilkinson et al., 2003), but even the most highly inlinked pages (in UK university web sites) rarely contain high quality scholarly content (Thelwall, 2002e).

Mathematical models of hyperlinking have been used to identify a significant geographical relationship for university web site interlinking (Thelwall, 2002b): closer universities tend to interlink more than distant universities, after taking the source and target institution research productivity into account (Thelwall, 2002a). Geographic factors have not been taken into account in subsequent mathematical modelling, however.

Although most previous research has focused on individual nations, there have also been a few international comparative studies of university web sites in the European Union (Thelwall et al., 2002) and in the Asia-Pacific region (Thelwall & Smith, 2002) but these have not included new mathematical models.

Research has also been carried out on inter-departmental linking within university web sites and found that counts of links to departments are associated with the research reputation of these individual departments (Chu, He, & Thelwall, 2002; Li, Thelwall, Musgrove, & Wilkinson, 2003; Tang & Thelwall, 2003), mirroring previous university-level findings. Other departmental-level research has produced diagrams of the links between sets of departments (Chen, Newman, Newnam, & Rada, 1998) and sought reasons for inter-subject linking (Thelwall, Harries, & Wilkinson, 2003), without creating new link models.

In summary, significant relationships and patterns can be extracted from the analysis of web links between academic institutions. It is now becoming accepted that it is important for researchers, and the academic community in general, to benefit from mining the linking of scholarly materials on the web in various ways (e.g., Jepsen, Seiden, Ingwersen, & Björneborn, 2004). Nevertheless, the question of how best to extract patterns from web links remains problematic in some respects. Previous studies of academic web interlinking have tended to hypothesise that the relationship between the research of a university and links to its web site should follow a linear trend, yet the typical distribution of web data, in general, seems to be a non-linear power law. The aim of this paper is to assess whether a linear trend or a power law is the most appropriate method with which to model the relationship between academic web size, interlinking data and research productivity indicators, continuing previous work exploring a range of informetric properties of academic web data (Payne & Thelwall, 2004). If the academic web were shown to demonstrate power law behaviour, it would have implications for previous web metric based research, because specially modified 'scale-independent' indicators would then need to be developed (Katz, 2000).

2. Previous related research

2.1. Web impact factors

Basing research on the concept of informative comparisons and drawing extensively upon the traditions of bibliometrics, Ingwersen (1998) introduced the web impact factor (WIF), a metric designed to assess the impact of an area of the web based upon counting the number of pages (or other units, as described below) providing hyperlinks to it. The WIF has also made a significant contribution to academic hyperlink research by enabling deeper mining of university web link data through the factoring out of the research productivity-related component (Li, 2003).

The WIF has since proved to be a popular tool for information scientists analysing web links because WIF variants, despite measuring Internet phenomena, produce results that are susceptible to a wider interpretation (Thelwall, 2002c). For example, Smith (1999) found the WIF a “useful measure of the overall influence of the web space”, when specifically discussing universities and research institutions. Two new web link count metrics have also been introduced, the web usage factor and the web connectivity factor, which are based upon site outlinks and the links between pairs of sites, respectively (Thelwall, 2003).

Most research into academic web interlinking using the WIF has tended to assume that web data should follow a linear trend (Thelwall, 2002a, 2002d; Thelwall & Wilkinson, 2003a). However, if academic web data is shown to demonstrate power law behaviour, it will affect the design of web metrics such as the WIF, and calculations based on web metrics would have to be modified to take this into account.

2.2. Power law research

The research of Adamic and Huberman (1999) has found that a power law was at work in the distribution of web pages so that a diminishingly small proportion of sites had an increasingly large page count. (A power law occurs when the frequencies n of some variable x , are proportional to $1/x^n$.) Their studies found that web growth dynamics and topology follow a power law distribution, a type of ‘rich get richer’ phenomenon: when new links are added to the web, they tend to connect pages that are already highly connected. This means that the web does not follow the usual mathematical models of random networks, but instead exhibits the type of physical order found in magnetic fields, galaxies and plant growth. Subsequent research showed that many other phenomena related to web links also follow power laws (Broder et al., 2000; Pennock, Flake, Lawrence, Glover, & Giles, 2002). The existence of a power law in the growth of the web not only implies the lack of any natural scale for the web, but also allows the expected number of sites of any given size to be determined without exhaustively crawling the web.

There are many other examples of power law distributions to be found within the confines of the Internet, including the number of intermediate links along link paths between web pages (Albert, Jeong, & Barabási, 1999). Rousseau (1997) found a power law for distribution of top-level domains and links counts from AltaVista to a topic-based set of pages. Barabasi, Albert, and Jeong (1999) showed that power laws apply to vertex connectivities in many large networks including the web, while Faloutsos (1999) discovered three power laws that characterised the Internet router topology on an inter-domain level during the year 1998. Power laws, albeit with anomalies, have also been discovered within the graph structures of the national university publicly indexable webs from Australia, New Zealand and the UK (Thelwall & Wilkinson, 2003b).

Although the connectivity distribution over the entire web is close to a pure power law, the distribution within specific categories is typically unimodal on a log scale, with the location of the mode, and thus the extent of the ‘rich get richer’ phenomenon, varying across the differing categories (Pennock et al., 2002). There seems to be a second linking law working in parallel with the power law, one that distributes links at random rather than predominantly to pages that are already highly linked. A consequence of this is that among web pages of the same genre (e.g. university home pages), the body of the distribution of inbound

links deviates strongly from a power law, exhibiting a roughly log-normal shape. The revised model combines a power law with random linking. This made it easier to incorporate the possibility for new sites to gain high link counts. This model also explained the partial non-linearity of many power graphs, as a slightly hooked shape is evident, albeit to differing degrees, at the top left of each graph. The hooked shape in the graph indicates that this latter tendency is present, but is still dominated by preferential attachment.

Power law models have become increasingly sophisticated, incorporating additional variables. For example, [Adamic and Huberman \(2002\)](#) have shown that site size, like site traffic, is highly skewed: a small number of sites are responsible for a disproportionately large part of the web's volume and activity. In their model, the growth of the web is subject to two dynamics: the fact that the total number of sites is growing exponentially, and the fact that the fluctuations in the size of a particular site are proportional to the size of the site, "The more pages a site has, the more likely it is that more pages will be added to it, just like the growth of a tree." And, like a tree, they hypothesised that the total size of the web will eventually become subject to resource limitations, but that at the moment, the current web is still just a sapling, with plenty of potential for continued exponential growth.

2.3. *Link analysis applications*

The themes that emerge from link analysis research are valuable in a number of respects. Analysis of the link structure of the web suggests that the on-going process of page creation and linkage, while very difficult to understand at a local level, results in structure with a degree of order. Thus, it gives a global understanding of the ways in which independent users build connections to one another in hypermedia that arises in a distributed fashion, and it provides a basis for predicting the way in which on-line communities will develop as they become increasingly connected (see [Björneborn, 2004](#)). Web based data, if sufficiently understood, have the potential to produce a new generation of science and technology ([Aguillo, 1998](#)) or financial ([Uberti & Maggioni, 2004](#)) indicators for the digital era, as well as supporting a range of social science research (e.g., [Garrido & Halavais, 2003](#); [Park, Barnett, & Nam, 2002](#)).

Some research has been carried out into the informetric properties of the Internet following claims that search engine tools are not sufficiently developed to produce accurate results ([Gibson, Kleinberg, & Raghavan, 1998](#)). Much of this type of research has been driven by the need to improve the quality and reliability of the variety of harvesting and ranking tools currently available to access information on the web (e.g., [Brin & Page, 1998](#); [Henzinger, 2001](#)).

3. Research question

Previous research into academic web interlinking has tended to hypothesise that the relationship between links to a university and its research productivity should follow a linear trend ([Thelwall, 2002a, 2002d](#); [Thelwall & Wilkinson, 2003a](#)). Yet the distribution of links to individual academic pages follows a power law ([Thelwall & Wilkinson, 2003b](#)). What has not been conclusively demonstrated, however, is that the relationship between link counts, when aggregated at the university level, and research productivity follows a linear rather than a non-linear power law. Put another way, should the links that a university creates or attracts always be proportional to its research productivity, or should larger universities expect to attract or create more (or less) links than this, by virtue of their size?

If the link count-research productivity relationship does follow a power law, then this has important implications for the design of web metrics such as the web impact factor ([Ingwersen, 1998](#)). For example, [Katz \(2000\)](#) has shown that specially modified 'scale-independent' indicators need to be developed for data that obeys a power law (see also [Leydesdorff & Bensman, 2005](#)). If power law behaviour were to be exhibited, then calculations based on WIFs would have to be modified to take this into account.

The primary research question is therefore as follows: *is the relationship between academic web and research productivity indicators best modelled by a linear trend or a power law?* In order to answer this question, this paper will concentrate on the following more specific sub-questions: *for UK university web sites, is the relationship between site size and research productivity, and between outlinks and research productivity best modelled by a linear trend or a non-linear power law?*

4. Methods

4.1. Raw data

The raw link data used during this study was derived from the text files of 111 UK universities as of July 2002, the link structure of which was obtained from a publicly available indexable database (Database 9) as part of the Wolverhampton University Academic Web Link Database Project (Thelwall, 2002/3). The database represents not the entire academic website in each case but only those pages that can be found by following links from the home page, excluding recognised mirror sites and pages from which robots are banned. In the case of universities with non-HTML links on the home page, the starting point chosen was a list of departmental home pages instead. The database does not include any identified areas of websites that are mirror copies of documentation produced elsewhere, although this process is error-prone due to its reliance upon human intervention.

The database used was created by a specialist information science web crawler (Thelwall, 2001b, 2004b). Essentially it crawls all HTML pages on a site that it can find by following links, but discards duplicate pages and bypasses pages that are identified by the webmaster as off limits to crawlers. This crawler covers websites accurately in the sense of comprehensively testing for and eliminating duplicates but the results cannot claim to be complete because of the robots.txt convention denying access to some sites and the crawler only being able to find pages by following HTML format links from other known pages. A web crawler is preferable to the use of a commercial search engine for the same statistics because of the greater degree of control by the researcher over the data collection (Thelwall, 2004b).

The UK academic web was surveyed in order to obtain statistics about web pages that were the target of links from web pages on other UK academic websites. Each university has a website with a home page accessible through at least one domain of the form www.name.ac.uk, where 'name' is an abbreviation of the university name, for example www.wlv.ac.uk for Wolverhampton university (Thelwall, 2001a).

The resultant link structure database consists of a separate text file for each university, giving a list of the URLs of all source pages crawled together with all identified target URLs referred to in the page, with duplicate URLs removed and all URLs truncated at the first '#' character (Thelwall, 2002a). This last point means that in one page there cannot be links to two or more parts of a common target page.

The university text files were processed by a specially written program which sorted the link data, removed all duplicates and aggregated the resultant data, i.e. source URL counts and target URL (link) counts from each university crawl, using alternative document models (ADMs). The basic data sets are therefore for each site (a) the set of all unique URLs of pages crawled, and (b) the set of all outlinks found in all pages crawled. Note that site self-links are included in the data set.

4.2. Alternative document models for the web

Although the web page is the typical indivisible unit of information in search engine results, is not necessarily the ideal primary source document for counting purposes (Thelwall, 2002a). When web links are the object of study, other alternatives, notably alternative document models involving the aggregation of pages based on domains and directories, have the potential to produce better results.

The original ADMs exploited simple URL-based heuristics to automatically merge web pages for counting purposes (Thelwall, 2002a). The different levels of aggregation of pages and links were an attempt to remove the effect of spurious duplication of links which may occur when a website contains identical links on many pages.

A second approach uses the same ADMs but ignores repeated links from one university to the same target web document at another university (Thelwall & Wilkinson, 2003a), attempting to approximate the range of Internet collaboration rather than its intensity. ADM metrics at different levels of aggregation using this approach effectively assess the range of the connection to each target university, although the outcomes are similar to the previous metrics.

It has been discovered that the domain and directory models were able to successfully reduce the impact of anomalous linking behaviour between pairs of websites, with the directory-based URL counting model being better for analysing interlinking between universities, at least in the UK (Payne & Thelwall, 2004; Thelwall & Wilkinson, 2003a). However, the domain ADM is also a good choice, for example to compare links between individual UK, Australian and New Zealand universities (Thelwall, 2004a). Given the lack of a universal best choice of ADM, the best option seems to be to use several (e.g., Thelwall & Vaughan, 2004).

The ADMs used within this paper aggregate pages at the web page, directory and domain level using the following definitions:

Page: Each separate HTML file is treated as a document for the purposes of extracting links. Each unique link URL is treated as pointing to a separate document for the purposes of finding link targets. A web page in this context is identified with its URL. Any URL starting with http:// is allowed and URLs will be truncated before any internal target designator symbol to avoid multiple links to different parts of the same page.

Directory: All HTML files in the same directory, for both sources and targets, are treated as a document. All URLs are automatically shortened to the position of the last slash, and links from multiple pages in the same directory are combined and duplicates eliminated.

Domain: As above except all HTML files with the same domain name are treated as a single document. This clusters together all pages hosted by a single subdomain of a university site. Domains are obtained by stripping any directory structure, file name, port number and password information from URLs, i.e. truncating each target URL just before its first slash, if present.

4.3. Staff numbers and RAE ratings

Fundamental to the research question is the definition of ‘academic web size’. Previous studies have used a number of factors in order to describe university site size, including the number of web pages within the university source domain, the number of university staff and the physical size of the university.

Following on from the work of Ingwersen (1998) in creating the external relative WIF (a count of all pages outside a site that contain a link to it, divided by the number of pages in this target site), a version of this external relative WIF was created especially for academic web sites (Thelwall, 2001c). This uses the number of full-time academic members of staff at the university instead of a site page count as the denominator of the calculation.

WIF metrics using this calculation, together with UK Research Assessment Exercise (RAE) derived data have been previously utilised revealing statistically significant correlations (Li et al., 2003; Smith & Thelwall, 2002; Thelwall, 2002c, 2003). Since the subject of the research question in the present paper has important implications for the design of web site metrics, and also as it is the measure used in most of the linear trend-based papers employing UK link data, like the present one, it seems natural to use the definition of research productivity as academic staff numbers multiplied by the average RAE score for each individual university. This is used in preference to web site page counts, as site page counts have proved to provide unreliable data (Thelwall, 2001a).

The average RAE rating of the universities was taken from the Times Higher Education Supplement (Mayfield University Consultants, 2001), which averages the grades awarded to each university by the Government Research Assessment Exercise. This is a peer review, subject-based process that is used to direct Government research funding. Staff numbers were taken from the Noble Publishing Co. (1999). The combination of data from different years is not ideal, although staff numbers and RAE averages are relatively stable for most universities and so should not significantly impact upon the results.

4.4. Statistical analysis

The statistical analysis program SPSS 10.0 was used to perform a number of tests on both the raw and ADM aggregated data, including non-parametric testing, linear regression and correlation analysis.

Linear regression analyses the relationship between two variables, x and y . For each subject, both x and y are known, and it is required to find the best straight line through the data. In general, the aim of linear regression is to adjust the values of slope and intercept to find the line that best predicts y from x . More precisely, the goal of the regression is to minimize the sum of the squares of the vertical distances of the points from the line. Note that linear regression does not test whether the data is linear. It assumes that the data is linear, and finds the slope and intercept that make a straight line best fit the data.

The equation of the power type trend line displayed in logarithmic graphs takes the form $y = cx^a$. That is, y is a power law in x with a power or index of a multiplied by a number or normalisation constant c .

Now, if it can be shown that the linear power 1 is in the 95% confidence interval for the power law parameter for the equation index, this can be used to claim that there is no evidence of a power law. This is because if $y = cx^a$, where $a = 1$, then $y = cx$, which is a linear equation, and this would mean that the relationship between UK university link data and research productivity indicators would be best modelled by a linear trend, and not a power law.

It is worth noting that there are direct techniques for testing for power laws available, and there is software to easily do this (Rousseau & Rousseau, 2000). However, the current paper does not directly test for a power law, as a power law fit could also be a linear trend fit and so would not answer the primary research question.

5. Results

The 111 raw data university text files were analysed and aggregated into directories and domains using the alternative document models, and statistics produced for source and target domains, directories and pages for each individual university. These were taken as dependent variables and plotted on six scatter graphs against the 'staff number * RAE rating' data as independent variables. All variables were subjected to the Kolmogorov–Smirnov test, the principal goodness of fit test for normal and uniform data sets, and the test distribution for each was found to be not significantly non-normal.

Figs. 1–6 show the data on a linear scale and Figs. 7–12 show the same data on a logarithmic scale.

Visual inspection of the graphs suggests that linear trends and power laws both fit the data sets to a reasonable extent, although it is also clear that the data does not exhibit the very marked power law pattern seen in previous web exercises, in which the power law slope is in the opposite direction to the above graphs and shows a marked 'broomstick' or 'hooked broomstick' shape, reflecting a 'rich get richer' phenomena. The graphs above resemble more those of Katz (2000), pointing to the possibility that there are economies of scale operating. However, to prove whether or not the best fit for these graphs is a linear or logarithmic line, more rigorous tests need to be applied. Each data set was thus subjected to linear regression techniques using SPSS, after a logarithmic transformation had been performed on both sets of variables.

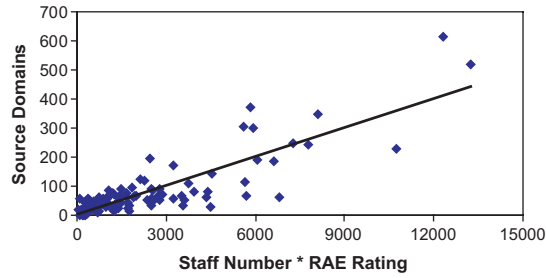


Fig. 1. Source size (number of unique domain names in each site) against research (staff number * RAE rating).

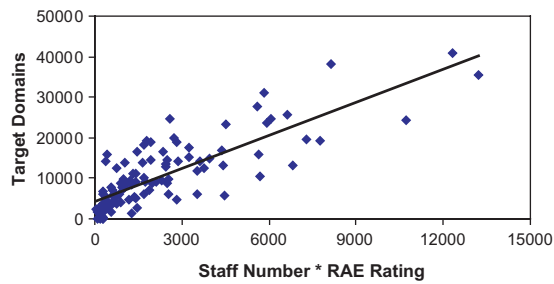


Fig. 2. Target size (number of unique domains targeted by links from each site) against research (staff number * RAE rating).

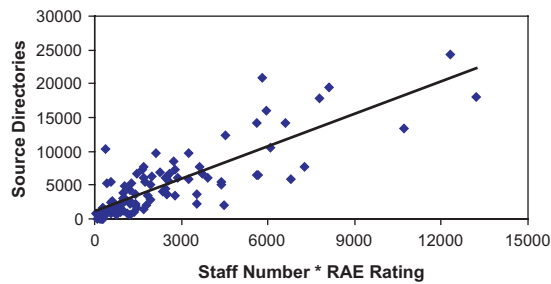


Fig. 3. Source size (number of unique directories in each site) against research (staff number * RAE rating).

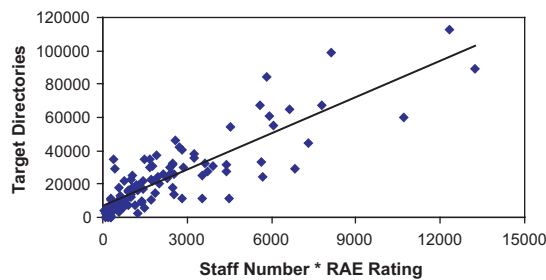


Fig. 4. Target size (number of unique directories targeted by links from each site) against research (staff number * RAE rating).

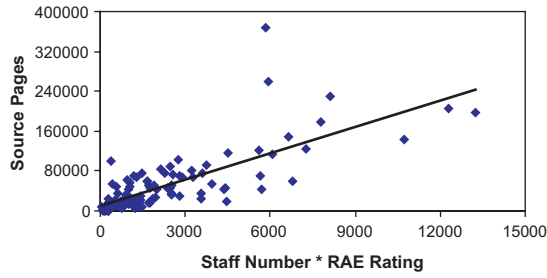


Fig. 5. Source size (number of unique source pages in each site) against research (staff number * RAE rating).

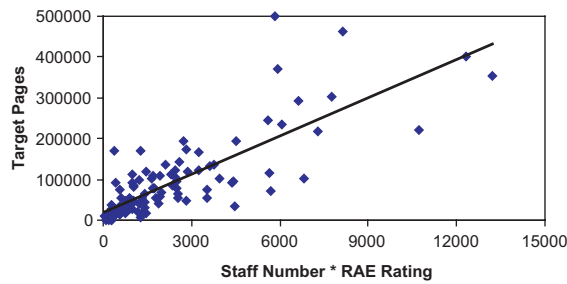


Fig. 6. Target size (number of unique pages targeted by links from each site) against research (staff number * RAE rating).

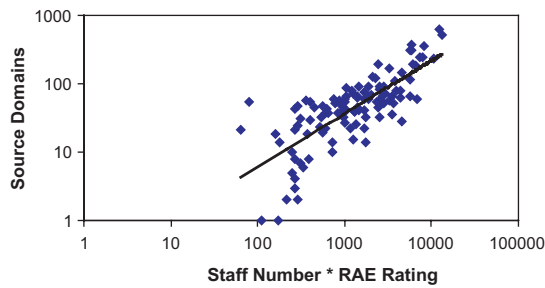


Fig. 7. Source size (number of unique source domains in each site) against research (staff number * RAE rating)—logarithmic graph.

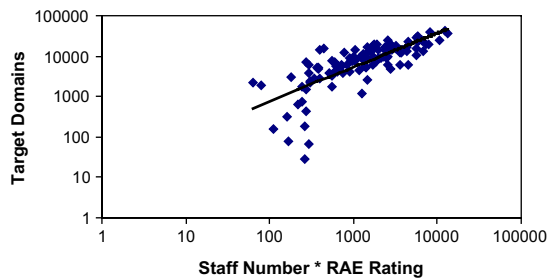


Fig. 8. Target size (number of unique domains targeted by links from each site) against research (staff number * RAE rating)—logarithmic graph.

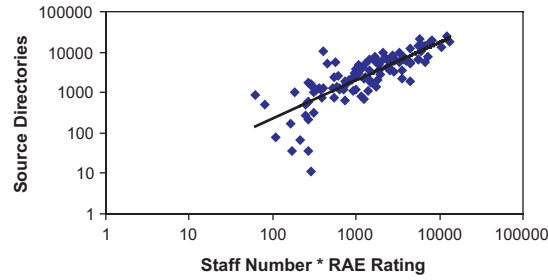


Fig. 9. Source size (number of unique source directories in each site) against research (staff number * RAE rating)—logarithmic graph.

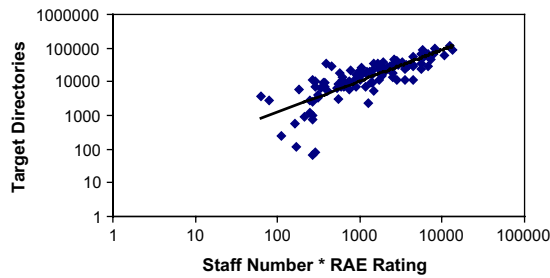


Fig. 10. Target size (number of unique directories targeted by links from each site) against research (staff number * RAE rating)—logarithmic graph.

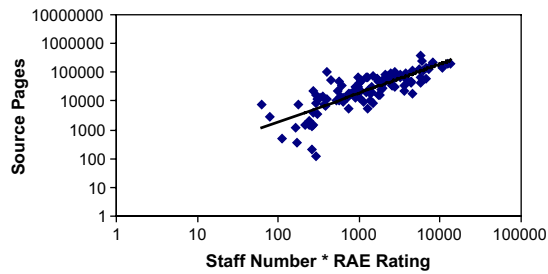


Fig. 11. Source size (number of unique source pages) against research (staff number * RAE rating)—logarithmic graph.

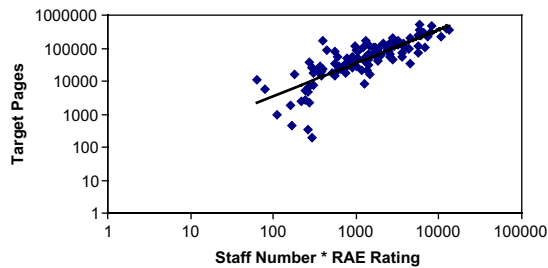


Fig. 12. Target size (number of unique pages targeted by links from each site) against research (staff number * RAE rating)—logarithmic graph.

95% confidence intervals were taken for the power in the power laws for each data set. This means that, given the assumptions of linear regression, the confidence interval is 95% sure to contain the best-fit regression line, leaving a 5% chance that the true line is outside those boundaries. However, this is not the same as saying that it will contain 95% of the data points. The results for each of the logarithmic graphs are shown in Table 1 below.

Table 2 gives Pearson's correlation coefficient, R , which gives a measure of linear association.

All graphs display correlation coefficients in a very narrow range between 0.742 and 0.794. Levels of correlation of between 0.7 and 0.9 can be described as high and so, using Pearson's correlation coefficient, R , it is clear that the graphs displayed in Figs. 7–12 all display high levels of correlation, i.e. significant at the 0.01 level (two-tailed).

6. Discussion

6.1. Power law or linear trend?

It can be observed from Table 1 that the linear power 1 does lie within the 95% confidence interval for the graphs of source directories vs. staff number * RAE rating, target directories vs. staff number * RAE rating, source pages vs. staff number * RAE rating and target pages vs. staff number * RAE rating. So it could be said that, for these graphs, it would be reasonable to use a linear model for the relationship. However, the linear power 1 does not lie within the 95% confidence interval for the graphs of staff * RAE vs. source domains and staff * RAE vs. target domains, although upper 95% confidence intervals of 0.901 and 0.986 respectively suggest that there may be evidence of linear trends which may be masked by inaccuracies in the data, inadvertently introduced as a result of the limitations highlighted in Section 6.3.

The high levels of correlation shown in the results, significant at the $p = 0.01$ level (two-tailed), do seem to suggest a definite linear relationship between the staff number * RAE rating data and the number of source domains, target domains source directories, target directories, source pages and target pages, and go some way in justifying the choice of methodology used.

It should be noted that all six logarithmic graphs show the same basic pattern with the two graphs utilising the domain ADMs containing the most extreme outliers. These are also the graphs that show the most deviation from the 95% confidence intervals and the lowest levels of correlation.

Table 1
Upper and lower 95% confidence intervals for power law powers (the line slopes in Figs. 7–12)

| ADM | Source documents | Target documents |
|-----------|------------------|------------------|
| Domain | (0.652, 0.901) | (0.698, 0.986) |
| Directory | (0.798, 1.077) | (0.786, 1.074) |
| Page | (0.863, 1.162) | (0.863, 1.157) |

Table 2
Pearson's correlation coefficient for Figs. 7–12

| ADM | Source documents | Target documents |
|-----------|------------------|------------------|
| Domain | 0.764 | 0.742 |
| Directory | 0.787 | 0.776 |
| Page | 0.789 | 0.794 |

6.2. Outliers

The following universities were noted as reoccurring outliers for the corresponding charts:

Paisley. This is not only the smallest university in terms of number of subdomains but also has the smallest number of source directories, source and target pages.

Thames Valley University (TVU). TVU has the lowest average RAE rating of 0.4, and this contributes to the lowest staff number * RAE rating.

Newport. This university has the joint second lowest RAE rating (0.5), and the second lowest staff number * RAE rating.

Chichester. Although not the smallest university, it does have the smallest number of both target domains and directories. The large number of source page links is explained by the design of its web page menu structure.

Because of the combination of low numbers of full-time academic staff and low average RAE scores, TVU and Newport have the lowest staff number * RAE rating (63.04 and 80.15 respectively) by some margin—the next lowest being Anglia with 361.1. The removal of the outliers with the lowest RAE ratings, i.e. Newport and TVU universities, brings the 95% confidence interval for the target domains vs. staff number * RAE rating graph to 1.048 and the 99% confidence interval for the source domains vs. staff number * RAE rating graph to 1.026, i.e. the linear power 1 is now within the respective confidence intervals for both of these graphs. The removal of outliers could be considered to be a permissible step in statistical analysis, and in this case, does not change the impact of the conclusions.

6.3. Limitations

There are a number of limitations of this study, some of which are highlighted in the list below:

- Universities have different policies for domain name use (Thelwall & Harries, 2003). This may explain the deviation in the source and target domain 95% confidence intervals.
- The study only covers one national university system, crawled at one time. It is possible that domain structure in other countries would be different, and so the results would not necessarily extend. Future results confirming the results for other countries would strengthen the findings.
- The study covers only web site size and outlink counts. Although it seems unlikely, it is possible that different results could be obtained from site inlink counts, which are most commonly used in webometrics, as well as site outlink counts (i.e. excluding site self-links), and the interlinking counts used in the WCF calculation.
- The validity of the non-web data used is not beyond question. The RAE rating figures used are now a number of years old, and the methods of counting academic staff numbers may vary considerably between universities.
- The staff numbers, RAE ratings and web link data were derived from three different years (1999, 2001 and 2002 respectively). While no significant change in the data is expected, minor inaccuracies may be introduced as a result of this time-span.
- There may be inaccuracies within the hyperlink structure web data such as typographical errors in target URLs.
- The web crawler used to collect the web data covers only the publicly indexable pages on the sites covered.

These are clearly drawbacks that should encourage caution in the interpretation of the conclusions as these limitations, some of which would be inherent to any similar web site counting study, make it much harder to identify mathematical patterns, even when they are actually present. It is hoped that future

web link research will improve the reliability of the data and increase confidence in the results but until then, the evidence presented here is supportive of the hypothesis that the relationship between academic web size and research productivity indicators is best modelled by linear trends.

7. Conclusions

This paper has taken as raw data the text files of UK university hyperlink structure produced by the Wolverhampton University's Statistical Cybermetrics Research Group, aggregated them into source and target domain, directory and page alternative document models, and displayed the results against academic staff numbers multiplied by average RAE rating data on both linear and logarithmic graphs in an attempt to show whether a linear trend or a power law is the most appropriate method with which to model the relationship between web publishing/linking and research productivity.

Following linear regression, analysis of the logarithmic graphs showed that the linear power 1 fell within the 95% confidence interval for the following graphs:

- Number of unique source pages vs. staff number * RAE rating.
- Number of target pages vs. staff number * RAE rating.
- Number of unique source directories vs. staff number * RAE rating.
- Number of target directories vs. staff number * RAE rating.

It is a reasonable interpretation then, that this data, using the page and directory ADM, is broadly consistent with a linear trend. The results using the domain ADM were not as conclusive, although even with no further analysis, with upper 95% confidence intervals of 0.901 and 0.986 for source domain ADM and target domain ADM respectively, it could be claimed that a linear trend is still in evidence. The deviation could be explained by inaccuracies in the raw data collection method, application of the ADM models, use of the staff number/RAE data or a combination of all three.

The situation for the domain ADM is different. There was some evidence of a non-linear power law, even after the removal of prominent outliers. The linear power 1 fell outside of the 95% confidence interval, even though it was within the 99% confidence interval. Inspection of the graph, however, does not show a clear power law trend; the clearest non-linear factor is that low research productivity universities have few domains. A non-linear power law would explain this, but an alternative explanation would be a break-point: perhaps universities need a critical research size in order to create the infrastructure (technical knowledge, and authority permission) to be able to create new subdomains as and when needed. In contrast to domains, directories typically do not need specialist knowledge, permission and access to create and so are in a sense more natural. This argument supports the use of the directory ADM as preferential to the domain ADM, at least until this apparent anomaly in domain creation is explained or disappears.

The fact that the relationship between academic web site size and research productivity is best modelled by a linear trend validates the results of previous research into academic web interlinking which has assumed that web data should follow a linear trend (Thelwall, 2002a, 2002d; Thelwall & Wilkinson, 2003a). It also justifies the design and use of web metrics such as the web impact factor (Ingwersen, 1998), which rely on linear data. Katz (2000) has shown that specially modified 'scale-independent' indicators need to be developed for data that obeys a power law. However, many of the 'rich gets richer' type of power law behaviour observed within the web concentrate on the distribution of overall web links over time, and is not strictly applicable to the collection of academic text file hyperlink structures used within this research, as they represent a snapshot of UK University hyperlink data as of July 2002, although this is the same raw data used in many previous studies of academic web links.

The results from this study hold importance as a step towards understanding the phenomenon of academic web linking and developing metrics to extract useful information. A greater understanding of the mathematical patterns and relationships within the hyperlink structure of these links will develop an appreciation of the way the academic web is presently connected, and may prove to be a useful tool in predicting future development and evolution.

It is evident that academic hyperlink analysis can be complex and problematical. Although significant mathematical patterns can be extracted from hyperlinks, it is still the case that they are a largely unregulated phenomenon. As a result great care must be taken to validate data when conducting hyperlink analyses to avoid drawing false conclusions because of data unreliability. Nevertheless, these positive results strengthen the case for using web link analysis as a tool with the potential to reveal underlying trends in academic web site interlinking.

Despite the problems mentioned above, the confidence intervals for the respective ADMs vs. staff number * RAE rating data are statistically significant. The positive results can confirm that it is a linear trend, and not a power law, which best models academic web size and interlinking data.

References

- Adamic, L. A., & Huberman, B. A. (1999). Internet: Growth dynamics of the World Wide Web. *Nature*, 401, 131.
- Adamic, L. A., & Huberman, B. A. (2002). Zipf's Law and the Internet. *Glottometrics*, 3, 143–150.
- Aguillo, I. F. (1998). STM information on the Web and the development of new Internet R&D databases and indicators. *Online information 98: Proceedings*, pp. 239–243.
- Albert, R., Jeong, H., & Barabási, A. L. (1999). The diameter of the World-Wide Web. *Nature*, 401, 130.
- Almind, T. C., & Ingwersen, P. (1997). Informetric analyses on the World Wide Web: Methodological approaches to 'webometrics'. *Journal of Documentation*, 53(4), 404–426.
- Barabasi, A., Albert, R., & Jeong, H. (1999). Mean-field theory for scale-free random networks. *Physica*, 272, 173–187.
- Bar-Ilan, J. (2001). Data collection methods on the Web for informetric purposes: A review and analysis. *Scientometrics*, 50(1), 7–32.
- Bar-Ilan (2004). Self-linking and self-linked rates of academic institutions on the web. *Scientometrics*, 59(1), 29–41.
- Björneborn, L. (2004). *Small-world link structures across an academic Web space: A library and information science approach*. PhD Thesis, Royal School of Library and Information Science, Copenhagen, Denmark.
- Björneborn, L., & Ingwersen, P. (2001). Perspectives of webometrics. *Scientometrics*, 50(1), 65–82.
- Björneborn, L., & Ingwersen, P. (2004). Toward a basic framework for Webometrics. *Journal of the American Society for Information Science and Technology*, 55(14), 1216–1227.
- Brin, S., & Page, L. (1998). The anatomy of a large scale hypertextual Web search engine. *Computer Networks and ISDN Systems*, 30(1–7), 107–117.
- 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.
- Chen, C., Newman, J., Newnam, R., & Rada, R. (1998). How did university departments interweave the Web: A study of connectivity and underlying factors. *Interacting with Computers*, 10, 353–373.
- Chu, H., He, S., & Thelwall, M. (2002). Library and Information Science Schools in Canada and USA: A Webometric perspective. *Journal of Education for Library and Information Science*, 43(2), 110–125.
- Cronin, B. (2001). Bibliometrics and beyond: Some thoughts on Web based citation analysis. *Journal of Information Science*, 27(1), 1–7.
- Egghe, L. (2000). New informetric aspects of the Internet: some reflections—many problems. *Journal of Information Science*, 26(5), 329–335.
- Faloutsos, M. (1999). On power law relationships of the Internet topology. In *ACM SIGCOMM 8/99*.
- Garrido, M., & Halavais, A. (2003). Mapping networks of support for the Zapatista Movement: Applying Social Network Analysis to study contemporary social movements. In M. McCaughey & M. Ayers (Eds.), *Cyberactivism: Online activism in theory and practice* (pp. 165–184). New York: Routledge.
- Gibson, D., Kleinberg, J., & Raghavan, P. (1998). Inferring Web communities from link topology. In *Hypertext 98: Ninth ACM Conference on Hypertext and Hypermedia, ACM, New York, USA*.
- Goodrum, A. A., McCain, K. W., Lawrence, S., & Giles, C. L. (2001). Scholarly publishing in the Internet age: A citation analysis of computer science literature. *Information Processing and Management*, 37(5), 661–676.
- Henzinger, M. (2001). Hyperlink analysis for the Web. *IEEE Internet Computing*, 5(1), 45–50.
- Ingwersen, P. (1998). The calculation of Web impact factors. *Journal of Documentation*, 54, 236–243.

- Jepsen, E. T., Seiden, P., Ingwersen, P., & Björneborn, L. (2004). Characteristics of scientific Web publications: Preliminary data gathering and analysis. *Journal of the American Society for Information Science and Technology*, 55(14), 1239–1249.
- Katz, J. (2000). Scale-independent indicators and research evaluation. *Science and Public Policy*, 27(1), 23–26.
- Larson, R. (1996). Bibliometrics of the World Wide Web: An exploratory analysis of the intellectual structure of Cyberspace. In *Proceedings of the 59th annual meeting of the American Society for Information Science* (pp. 71–78).
- Leydesdorff, L., & Benschman, S. (2005). Classification, powerlaws, and the logarithmic transformation. Available from <http://www.leydesdorff.net/log05/log05.pdf>.
- Li, X. (2003). A review of the development and application of the Web impact factor. *Online Information Review*, 27(6), 407–417.
- Li, X., Thelwall, M., Musgrove, P., & Wilkinson, D. (2003). The relationship between the WIFs or inlinks of Computer Science Departments in UK and their RAE ratings or research productivities in 2001. *Scientometrics*, 57(2), 239–255.
- Mayfield University Consultants (2001). League Tables 2001. *The Times Higher Education Supplement*. May 18, T2–T3.
- Noble Publishing Co. (1999). *Noble's higher education financial yearbook 1999*. Edinburgh: Noble Publishing.
- Park, H. W., Barnett, G. A., & Nam, I. (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.
- Payne, N., & Thelwall, M. (2004). A statistical analysis of UK academic Web links. *Cybermetrics*, 8(1), paper 2. Available from <<http://www.cindoc.csic.es/cybermetrics/articles/v8i1p2.html>> (accessed on 3 August 2004).
- Pennock, D., Flake, G., Lawrence, S., Glover, E., & 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(8), 5207–5211.
- Prime, C., Bassecoulard, E., & Zitt, M. (2002). Co-citations and co-sitations: A cautionary view on an analogy. *Scientometrics*, 54(2), 291–308.
- Rousseau, R. (1997). Situations: An exploratory study. *Cybermetrics*, 1(1), paper 1. Available from <<http://www.cindoc.csic.es/cybermetrics/articles/v1i1p1.html>> (accessed on 13 August 2004).
- Rousseau, B., & Rousseau, R. (2000). LOTKA: A program to fit a power law distribution to observed frequency data. *Cybermetrics*, 4(1), paper 4. Available from <<http://www.cindoc.csic.es/cybermetrics/articles/v4i1p4.html>> (accessed on 13 August 2004).
- Smith, A. G. (1999). A tale of two Web spaces: Comparing sites using Web impact factors. *Journal of Documentation*, 55(5), 577–592.
- Smith, A., & Thelwall, M. (2002). Web impact factors for Australasian universities. *Scientometrics*, 54(3), 363–380.
- Tang, R., & Thelwall, M. (2003). U.S. academic departmental Web-site interlinking in the United States Disciplinary differences. *Library and Information Science Research*, 25, 437–458.
- Thelwall, M. (2001a). Extracting macroscopic information from Web links. *Journal of the American Society for Information Science and Technology*, 52(13), 1157–1168.
- Thelwall, M. (2001b). A Web crawler design for data mining. *Journal of Information Science*, 27(5), 319–325.
- Thelwall, M. (2001c). Results from a Web impact factor crawler. *Journal of Documentation*, 57(2), 177–191.
- Thelwall, M. (2002a). Conceptualising documentation on the Web: An evaluation of different heuristic-based models for counting links between university Websites. *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 Website interlinking. *Journal of Documentation*, 58(5).
- Thelwall, M. (2002c). A comparison of sources of links for academic Web impact factor calculations. *Journal of Documentation*, 58(1), 66–78.
- Thelwall, M. (2002d). A research and institutional size based model for National University Website interlinking. *Journal of Documentation*, 58(6), 683–694.
- Thelwall, M. (2002e). The top 100 linked pages on UK university Web sites: High inlink counts are not usually directly associated with quality scholarly content. *Journal of Information Science*, 28(6), 485–491.
- Thelwall, M. (2002/3). A free database of University Web links data collection issues. *Cybermetrics*, 6/7. Available from <<http://www.cindoc.csic.es/cybermetrics/articles/v6i1p2.htm>> (accessed on 5 May 2004).
- Thelwall, M. (2003). Web use and peer interconnectivity metrics for academic Web sites. *Journal of Information Science*, 29(1), 1–10.
- Thelwall, M. (2004a). Methods for reporting on the targets of links from national systems of university Web sites. *Information Processing and Management*, 40(1), 125–144.
- Thelwall, M. (2004b). *Link analysis: An information science approach*. San Diego: Academic Press.
- Thelwall, M., Binns, R., Harries, G., Kennedy, T., Price, L., & Wilkinson, D. (2002). European Union associated university websites. *Scientometrics*, 53(1), 95–111.
- 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 better scholars' web publications have significantly higher 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., & Smith, A. (2002). Interlinking between Asia–Pacific university web sites. *Scientometrics*, 55(3), 363–376.
- Thelwall, M., & Tang, R. (2003). Disciplinary and linguistic considerations for academic Web linking: An exploratory hyperlink mediated study with Mainland China and Taiwan. *Scientometrics*, 58(1), 153–179.
- Thelwall, M., & Vaughan, L. (2004). New versions of PageRank employing alternative web documents models. *ASLIB Proceedings*, 56(1), 24–33.
- Thelwall, M., & Wilkinson, D. (2003a). Three target document range metrics for university Websites. *Journal of the American Society for Information Science and Technology*, 54(6), 489–496.
- Thelwall, M., & Wilkinson, D. (2003b). Graph structure in three national academic Webs: Power laws with anomalies. *Journal of the American Society for Information Science and Technology*, 54(8), 706–712.
- Uberti, T. E., & Maggioni, M. A. (2004). *Infrastrutture ICT e relazionalità potenziale. Un esercizio di “hyperlinks counting” a livello sub-nazionale*, Quaderno DISEIS n. 0402, Università Cattolica del sacro Cuore, Milano, Vita e Pensiero.
- van Raan, A. F. J. (2001). Bibliometrics and Internet: Some observations and expectations. *Scientometrics*, 50(1), 59–63.
- Vaughan, L., & Thelwall, M. (2005). A modeling approach to uncover hyperlink patterns: The case of Canadian universities. *Information Processing and Management*, 41(2), 347–359.
- 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), 49–56.