Methods for reporting on the targets of links from national systems of university Web sites*

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
Whilst hyperlinks within Web sites may be primarily created for navigation purposes, those between sites are a rich source of information about the content and use of the Web. As a result there is a need to derive descriptive statistics about them, both to help understand the underlying communication processes and so that policy makers can gain insights into the use of online information by those located within their constituency. It is known, however, that using the individual web link source page as the basic unit of counting is problematical because of the number and size of link anomalies. The challenge addressed in this paper is that of developing methods to assess techniques for counting links from groups of large university web sites (site outlinks). Two methods to assess the reliability of link counts are developed and applied to judge which of seven advanced document models are most appropriate in each case. The most generally applicable method used is an internal consistency test based upon a highly simplified model of web linking behaviour. The data used comes from crawls of UK, Australian and New Zealand universities. The standard domain advanced web document model emerges as the logical choice for comparison purposes within this set. Some descriptive statistics concerning Top Level Domain link targets are given and it is demonstrated that the choice of model can effect the final results.

Keywords: Web links; scholarly communication, linking models

* Information Processing and Management, to appear.

1. Introduction

1.1. Background

The issue of where Web sites link to is an important one for those who wish to trace scholarly activity and assess the patterns of information use in higher education. The Web is increasingly valuable as an information source for research (e.g. Kling & McKim, 2000; Koku *et al*., 2001; Lawrence, 2002), and there is a need to exploit the electronic trails of hyperlinks left by information providers within defined areas of the Web. In the context of university web sites patterns of informal scholarly communication may perhaps be traceable in a way that has not previously been possible (e.g. Cronin *et al*., 1998; Ingwersen, 1998; Goodrum *et al*., 2001) in order to complement existing bibliometric techniques (e.g. Glänzel, 2001). This may yield information about aspects of scholarly profiles that were previously difficult to gain concretely assess. One example of this is the use of an academic's research within university courses (Wilson, 2002). Cronin and Shaw (2002, in press) have also characterised web mentions as one potential measure of symbolic capital. Moreover, on a larger scale such information could be useful for managers to assess the value of Web access provision, for educators and librarians to identify information-rich sectors of the Web, for information scientists to study the process of information use and for policy makers to identify national and other emerging areas of strategic collaboration – or the lack thereof. The results could potentially also be used to give early warning of problems such as linguistic, cultural or technological barriers to information use. One example of this is the limited Web use in Eastern European universities being potentially an obstacle to integration into the research collaboration initiatives funded by the European Union (Thelwall *et al*., 2002). Potential general information mining approaches that could be used in these contexts include: benchmarking and identification of anomalous linking patterns within a coherent collection of sites; and comparisons between two or more groups of comparable sites, perhaps from different countries. Previous university-based link target analyses have mainly been restricted to the relatively tightly manageable target domains of sets of other university web sites, but it is also desirable to cast the net wider and investigate the full spectrum of link targets, and so investigate all sites outlinks. The object of this research will be to develop systematic methods to assess the extent and spread of link targets as categorized by URL Top Level Domains (TLDs). TLDs are relatively simple to extract from hyperlinks yet are collectively providers of potentially valuable information because of the formal, if imperfect, association of most with countries or organization types.

1.2. Related research

The strategic objective underlying this research is the obtaining of knowledge about patterns of Web use but hyperlinks are only one potential source of raw data for this. Two other useful repositories of information about web use are individual web server log files and search engine log files. The former can be usefully analysed in order to identify comprehensively patterns of use of resources hosted by a server (Nicholas *et al*., 1999). They can also be used to track how users arrived at the site in question (Thelwall, 2001a) but since they are not normally made publicly available, they are not useful as a general tool to track the use of large collections of web sites. Search engine logs are another potential source of information, particularly concerning information retrieval strategies and popular content categories (Spink *et al.*, 2001), but these are again not publicly available for inspection, although they could in theory be used for domain specific information about, say, which TLDs are visited most frequently by requests originating in the edu domain. As a result of the lack of freely available information sources on web use, there is a vacuum created by the need for general information about patterns of inter-site web use, one that the study of web links may partially fill.

Many indicators point to the valuable information that can be mined from web links although they are clearly an imperfect source of information about web use. Some modern

search engines such as Google operate on the basis that links contain valuable information that can be mined to improve the performance of page ranking algorithms (Brin & Page, 1998). In addition, one recently developed algorithm is able to use the link structure of the Web alone to identify self-organised communities (Flake *et al*., 2002), indicating that the semantics of the pages involved does not have to be used to extract information. On a larger scale it has been shown that aggregates of links between university Web sites can produce figures that correlate significantly with the target university productivity in the UK (Thelwall, 2001b), Australia (Smith & Thelwall, 2002), and to some extent China (Tang & Thelwall, 2002). A proposed research related model of web linking suggests that the same correlation should also be present for source university research productivity (Thelwall, 2002f). The correlations found should be treated with caution and do not prove a cause and effect relationship, however. A recent survey of a random collection of links between UK university Web sites found over 90% were related in some way to scholarly activities, but that less than one percent were equivalent to formal citations (Wilkinson *et al*., 2003). It was concluded that counts of academic Web links will be merging a range of predominantly academic factors but do not admit a simple rationalization. These findings suggest, however, that links may be a valuable potential source of information about scholarly use of the web. Whilst web authors are able to link to any pages almost at random, and some probably do (Stock, 2002), the evidence shows that there are enough patterns to make link analysis worthwhile if counts are made over a sufficiently large collection of pages to allow the law of averages to operate.

Some previous hyperlink research has focussed specifically on the pages targeted by a web site or system of web sites. Thelwall (2002a) identified the most highly targeted pages within UK universities, counting links only from other UK universities, and found that many were created to give credit to the target for some reason, rather than to indicate the location of useful information. A separate study on the PageRank algorithm (Thelwall, 2002b) included link counts and found that internal links tended to dominate external ones, reinforcing the view that in order to extract useful aggregate information, internal links should be excluded. A network diagram technique has been developed to produce a graphical representation of the links between large web spaces (Thelwall, 2001d; Thelwall & Smith, 2002) but the issue of assessing the robustness of simple link counts was not addressed.

Hyperlinks have also been the focus of studies from other unrelated perspectives, for example constructing derived maps in geography (e.g. Brunn & Dodge, 2001) and tracking online movements (Garrido & Halavais, 2002) or online business interconnectivity (Park *et al*., 2002) with social networks analysis.

Much of the previous webometric research into hyperlinks has used simple link counts, where links are counted based upon the full target URLs and without any attempt to remove duplicate links to the same target URL from different source pages (HTML files). This is a problematic approach because a web site can have repeated links on many pages, for example to a sponsoring organisation. To give a more extreme example, if an author posted one million pages on site A each with a link to a university in Brazil, then this would probably swamp the total link counts and it could be accurately reported that, "over 99% of links from site A are aimed at Brazil". The problem, then, is that the lack of quality control on the Web and the ease with which computer literate individuals can produce huge numbers of pages renders simple link count statistics potentially meaningless. As a result of this and other considerations, web based statistics must be treated with extreme caution (Egghe, 2000; Bar-Ilan, 2001; Björneborn & Ingwersen, 2001; van Raan, 2001). In response, however, new document models have been devised that aggregate links not on the individual file but on collections of files, defined in a total of seven different ways, described in the next section (Thelwall, 2002c; Thelwall & Harries, 2003; Thelwall & Wilkinson, 2003). These have proved very successful in the context of assessing links between universities within the same country, producing highly statistically significant correlations with measures of research productivity. For example, counting links between directories instead of between pages was shown to give total counts of links to each individual university that correlated better with research productivity for UK universities..

1.3. The Research Questions

Techniques to measure the consistency of alternative methodologies for the reporting on the extent of Top Level Domain targeting by links from national university systems will be developed and assessed. The seven advanced document models will be compared using these techniques, based upon data collected from crawls of the university web sites of Australia, New Zealand and the UK. The specific questions addressed are as follows.

- Are the results of the different techniques broadly consistent in recommending the same counting methodology?
- What kind of information can be gleaned by an analysis of TLD targets of national university systems?
- Do the different counting methodologies give significantly different results (i.e. is it worthwhile testing different ones to find the best)?

2. Methods for Counting Link Target Domains – Advanced Document Models

Before developing the techniques to assess the different counting methodologies, the seven major candidates will be briefly described (Thelwall, 2002c). The underlying advanced document models are based upon different levels of aggregation of both source and target web pages, each of which corresponds to an URL segmentation strategy. The counting methodologies that are based upon the models are described underneath.

- *Individual Web file/page* A file document corresponds to a single URL, after truncation before any internal target marker '#' character found to avoid multiple references to different parts of the same page. Typically a file document corresponds to a single electronic file on a server.
- *Directory* A directory document corresponds to a (partial) URL and encapsulates all URLs that are equal to it after truncation at their last slash. Typically, a directory document is a collection if files in the same directory on a server.
- *Domain* A domain document corresponds to a valid domain name, and encapsulates all URLs with that domain name. A domain document may correspond to a departmental web site, for example, if it has been given its own subdomain.
- *University* A university document corresponds to a canonical domain name registered for a university, including all subdomains and alternative equivalent domain names (e.g. wolverhampton.ac.uk = wlv.ac.uk).

The terminology of Björneborn (2001) will be adopted, where a *page outlink* is a link in a page that is targeted at a different page and a *page inlink* is a link to the page in question, also from a different page. Similar definitions apply to directory/domain/university/site inlinks and outlinks. A link is called an external link here if it is a university outlink. The seven counting methodologies all have different requirements for classifying links as duplicate. In the page/directory/domain/university counting methodology, two or more links from the same source page/directory/domain/university to the same target page/directory/domain/domain respectively, are considered to be duplicates and count collectively as only one. In the page/directory/domain range counting methodologies, multiple links to the same target page/directory/domain respectively from the same source university count as duplicates. This is equivalent to using the university model for link sources and the model above for link targets, and corresponds to a simple count of the number of different pages, directories or domains targeted by a university. A university range model would be the same as the standard university model, and so is not included.

3. Tests for the Best Fitting Document Model for TLD Target Counting

3.1. External Model Consistency Tests

In order to assess how effectively a model fits a collection of web pages, hypotheses must be made about the expected behaviour of counts of links, either in terms of associations with external measures or in terms of some property that can be internally calculated. In this section external measures are the focus. This approach has been previously used for counts of links *between* UK university web sites, using a measure of institutional research productivity as the external measure, described below in section 4.1. The results showed that the domain and directory standard and range versions were the best, with the university model using too great aggregation to give useful results, and the page model too little (Thelwall, 2002c; Thelwall & Harries, 2003; Thelwall & Wilkinson, 2003).

It is hypothesized that a research relationship may also be present for all university outlinks. The external consistency test based upon this assumption is that the appropriateness of a document model can be measured by the extent to which the results it produces correlate with the research productivity of the source institutions. If no significant association were to be found for any model then this would suggest that either the document models were all badly fitting, or that the test itself was inappropriate: i.e. that correlation with research is not a significant indicator of appropriateness of model. A practical problem with this, however, is that there are few known countries outside the UK for which there is a research assessment exercise sufficiently authoritative to be definitive for this purpose, although citation based indices may serve in places where reputable figures are not available or are only published in rank form.

It should be noted that the test, as described, is only applicable to collections of university web sites. For a different type of site it may possibly be the case that there is a different external measure of some aspect of the hosting organisations that it is appropriate to use instead of research productivity.

3.2. The ITTD Model of Linking Behaviour

In order to devise a test based only upon the internal structure of the data set, a theoretical model of linking must first be proposed. Without this, there is no basis from which judgements can be made. The model to be used will be described and then its fit with reality will be discussed. In defence of the general approach of applying simplistic models to web links, this tactic has been previously used with great success (e.g. Thelwall, 2002f; Pennock *et al*., 2002).

The Independent TLD Target Distribution (ITTD) model proposed here is based upon the simplifying assumption that hyperlink target TLDs display regular behaviour for similar sites within a country. This then allows investigations into the extent to which such regularity actually occurs in any observed data set. The target TLD is the most general information obtainable from the text of an URL without accessing other contents of its hosting page. National groups of similar sites are an appropriate choice for the scope of the model because each country has its own TLD and presumably links more extensively to this one than other countries would, indicating a significantly different TLD target pattern. The following simplistic assumptions are made about academic web linking patterns based upon these general principles. They can be applied to all of the document models described above, or indeed any others.

(1) Each university Web site is constructed from a finite collection of Web “documents”, however defined.
(2) Web documents from university sites within a given country come from a common probability distribution in terms of the likelihood of university outlinks by TLD.
(3) The outlinks for separate source documents from the probability distribution in (2) are statistically independent.

An example will be given to clarify the implications of the assumptions for a hypothetical country X. For simplicity the pages only contain links to one type of TLD, but this is not necessary in general.

Table 1
A probability distribution of document outlinks from country X.

Outlinks	Probability
1 to a com document	0.2
2 to a com document	0.1
3 to a com document	0.1
1 to an edu document	0.3
No outlinks	0.3

In a large sample of n documents from a university in country X, based upon the probability distribution in Table 1, it would be expected that there would be close to $0.3n$ edu document outlinks and $(1 \times 0.2 + 2 \times 0.1 + 3 \times 0.1)\, n = 0.7n$ com document outlinks.

Assuming that the model was perfect and that there was a sample of large university websites from a single country then some *ratios* calculated from the link figures for each university could be expected to be very similar. These would include the ratio of all links to a given well linked to TLD to all links. In the above example, universities in country X should give

$$\frac{\text{all edu site outlinks}}{\text{all site outlinks}} \approx \frac{0.7n}{n} = 0.7 \text{ and } \frac{\text{all com site outlinks}}{\text{all site outlinks}} \approx \frac{0.3n}{n} = 0.3$$

If the assumptions 1-3 were correct for a given country but a large variation in TLD ratios between universities in a country was discovered then it could logically only be inferred that *incorrect document models had been applied*. This, then, forms the theoretical basis for a test of internal consistency.

Of course assumptions 1 – 3 are all seriously flawed and will not apply to any national university system. Some important reasons why this is the case are given below.

(1) On the web, concepts of document and genre are very fluid and there is probably much web-based information that is in a format that could not be easily divided up into coherent documents from recognised genres (Crowston *et al*., 2000; Crowston & Williams, 2000; Nilan, *et al*., 2001; Rehm, 2002).

(2) Different types of purposes for web pages result in different profiles in the types of resources linked to. An academic web page may well be more likely to link to academic sources than one created by an administrator, perhaps. Similarly, the pages of researchers with a more applied focus presumably are more likely to link to .com domain sites than those of the more pure oriented. Thus it is not reasonable to believe that a common distribution would fit all.

(3) Factors such as social interactions by authors within a university and the creation of multiple documents by a single author are likely to result in clustering of links to individual sites and pages, breaking the independence assumption. It has been suggested that some types of links play the role of co-authorship attributions in academic papers (Thelwall, 2002g).

Despite these problems, if all the assumptions are approximately true when subject to averaging through application across a large university web site, then it is hypothesized that low fluctuations would be still be observed for reasonably well-fitting document models. Moreover, if an approximate model fit can be discovered then the deviations may become interesting sources of new information concerning anomalous behaviour of individual institutions. As an illustration of this, a rogue university with a different outlinking pattern may indicate an unusual underlying relationship to the Web or the international research community, such as a peculiarly national or international focus.

3.3. Internal Model Consistency Tests

Under the assumptions of the ITTD model, for a set of large university web sites within a nation and an effective document model heuristic the proportion of links to any commonly targeted top level domain would be expected to be approximately constant. The test for the internal consistency of a document model for a national university system is therefore that the variance in the proportion of counts from each university going to a top-level target domain is minimised. In addition, the proportions could be expected to be normally distributed. The TLD should be chosen to have as close as possible to 50% of all links to maximise the discriminatory power of this test. If the 'wrong' document model were to be applied, then a greater variety of proportions might be found either because too little aggregation left outliers in the data or because too much was obscuring a trend that included multiple documents that were being unnecessarily combined for counting purposes. The latter appears less likely, and so it is recommended that models with a lower degree of aggregation should be chosen over others that produce similar results.

It is not expected that a correction for degree of internal variance within a university Web site will be needed to compensate for the differing degrees of aggregation produced by the different document models. This would be an additional consideration, however, if the Web sites were small in terms of containing a relatively low number of outlinks under at least one of the models applied.

4. Experimental Method

4.1. Overall Approach

Internal and external consistency tests will be conducted for the UK university system. The assumption with the external test will be that good models will produce results that correlate highly with university research productivity. Such correlations have been shown to be present in inlinks from various generic TLDs to UK universities (Thelwall, 2002e). The main purpose of the suite of tests will be to see whether their results are consistent. The research productivity figures used are derived from the official UK 2001 government Research Assessment Exercise and essentially represent the average quality of research at a University times faculty size. See the official web sites (http://www.qaa.ac.uk, http://www.rae.ac.uk) for more information about the RAE and the *Education Guardian* (2001) for more on the average quality calculations.

A broader picture of the effectiveness of the models will then be assessed by applying the internal consistency tests to the national university systems of Australia and New Zealand. Together the three countries have a similar economic, linguistic and cultural background and so should provide a useful reference set for each other. If the results are positive then the variances for each methodology should indicate the best TLD counting model for each country, and the selection should be both clear cut and consistent across choice of TLD, whenever multiple choices are available. The fit of each document model will be further tested with normality tests on the ratio data. To merely identify the best document model counting methodology, only one test of significance would be needed, but the additional tests are included to assess the consistency of the consistency tests themselves.

4.2. The Data Sets

The raw data used comes from a crawl of 107 UK university web sites previously used from July, 2001 and two new crawls of Australian and New Zealand university sites, in December 2001 to January 2001, and January 2001 respectively. These come from a specialist information science web crawler, designed for accurate and comprehensive site crawling (Thelwall, 2001c, 2001e) but excluding, when identified, mirror sites such as copies of computer documentation. The crawl databases are all now publicly available (http://cybermetrics.wlv.ac.uk/database). Total external links extracted from these were 2,111,200 for Australia, 5,156,407 for the UK and 260,269 for New Zealand.

Four separate subsets were used for each test, representing all site outlinks, the national links and the two largest subsets by TLD.

- *Site outlinks* All links originating at a crawled page of a university and targeted at an URL that has a domain name not known to be associated with the source institution.
- *National links* All site outlinks originating at a crawled page of a university in a given country and targeted at an URL from the same top level domain (e.g. uk, au, nz).
- *edu/com links* All links originating at a crawled page of a university in a given country and targeted at an URL with a top level domain of edu/com, respectively.

The UK is probably unusual in the variety of its higher education institutions, and indeed it is government policy to promote diversity (Dearing, 1997). However, this means that there is an enormous variety in terms of both size and average research quality. As a result of this, it was decided to test two subsets of universities in addition to the whole set for the UK. The two subsets are the top half and bottom half from the perspective of total research productivity. The purpose of this is to be able to assess whether tests need a relatively homogeneous data source to be effective. The top half is more useful for this purpose since it is thought to be more similar to the set of research institutions that would be found in a developed nation than the bottom half would be.

5. Results

5.1. The UK 2001 Data Set

5.1.1. External Consistency Tests: Correlations Between Link Counts and Research Productivity

Spearman correlation coefficients were calculated for each model between counts of links from UK universities and research productivity. Spearman coefficients were used instead of Pearson because the distribution of research productivity figures was found to be significantly non-normal for all three data sets (Kolmogorov-Smirnov test).

- *All 107 universities* All correlations were highly significant and broadly similar in size, with the best model varying between subset so that no one model was clearly the best. Although links to the edu domain generally correlated best with research productivity, this does not help in the selection of document models.

Table 2
Spearman correlation coefficients for outlink counts compared to research productivity for all UK universities. Additional correlations are shown for four subsets of the data. All are significant at the 0.1% level. The 'best' model for each column is in bold.

	All site outlinks	National links	edu links	com links
File model	0.774	0.780	0.820	0.703
Directory model	0.804	**0.813**	0.825	0.744
Domain model	0.804	0.797	0.815	**0.759**
University model	0.782	0.710	0.801	0.739
File range model	**0.813**	0.799	**0.828**	0.747
Directory range model	0.797	0.794	0.814	0.749
Domain range model	0.793	0.763	0.804	0.740

- *The top 53 research output universities* Again all correlations are significant. They are generally lower than those for the complete set, reflecting the relative success in link counts differentiating between the lower and higher research rated institutions. Surprisingly, the standard file model performs quite well in the main category, as does the file range model.

Table 3
Spearman correlation coefficients for link counts compared to research productivity for all the top 53 research producing UK universities. All are significant at the 0.1% level. The 'best' model for each column is in bold.

	All site outlinks	National links	edu links	com links
File model	0.708	0.676	0.679	0.567
Directory model	0.674	0.685	0.710	0.594
Domain model	0.669	0.668	0.706	**0.639**
University model	0.628	0.560	0.716	0.596
File range model	**0.716**	**0.690**	0.714	0.597
Directory range model	0.666	0.658	**0.718**	0.599
Domain range model	0.649	0.613	0.713	0.599

- *The bottom 54 research output universities* Again all correlations are significant, but lower than those for the top 53, perhaps reflecting a smaller range of research productivities and therefore, less differentiation between institutions. This is a problem for the power of rank based tests. The standard directory and domain models are generally the best, and significantly better than the default file model.

Table 4
Spearman correlation coefficients for link counts compared to research productivity for all the bottom 54 research producing UK universities. All are significant at the 0.1% level. The 'best' model for each column is in bold.

	All site outlinks	National links	edu links	com links
File model	0.538	0.576	0.567	0.533
Directory model	0.615	**0.651**	0.546	0.565
Domain model	**0.620**	0.634	**0.602**	**0.567**
University model	0.597	0.581	0.562	0.545
File range model	0.599	0.587	0.565	0.548
Directory range	0.594	0.605	0.551	0.549
Domain range model	0.606	0.603	0.564	0.546

5.1.2. Internal consistency tests: standard deviations and normality tests for ratios of specified outlinks to all outlinks

- *All universities* The domain model is the best in both cases. Note that edu and com results are not consistent and that there are several significantly non-normal sets of proportions.

Table 5.
Normality tests (Kolmogorov-Smirnov p values are the top figures in the cells) and standard deviations (bottom) for UK universities. Italic cells are significantly non-normal and the bold cells have the smallest standard deviation in each column (n=109).

	edu links to all links ratio	com links to all links ratio
File model	*0.046* *0.0359*	0.137 0.119
Directory model	0.200+ 0.0328	*0.000* *0.0876*
Domain model	0.200+ **0.0324**	*0.004* ***0.0585***
University model	*0.010* *0.0351*	*0.018* *0.0640*
File range model	0.200+ 0.0386	*0.002* *0.0943*
Directory range model	0.200+ 0.0373	*0.000* *0.0808*
Domain range model	0.077 0.0339	*0.004* *0.0603*

- *Top 53 research output universities* These results are similar to those for all universities, but there is less evidence of non-normal behaviour.

Table 6.
Normality tests (top figure in each cell) and standard deviations (bottom) for the top 53 research output universities. Italic cells are significantly non-normal and the bold cells have the smallest standard deviation in each column.

	edu links to all links ratio	com links to all links ratio
File model	0.057 0.0353	0.104 0.0912
Directory model	0.200+ 0.0283	0.200+ 0.0463
Domain model	0.200+ **0.0270**	*0.030* ***0.0437***
University model	*0.039* *0.0307*	0.200+ 0.0479
File range model	0.200+ 0.0308	0.200+ 0.0579
Directory range model	0.200+ 0.0303	0.200+ 0.0453
Domain range model	0.084 0.0287	0.200+ 0.0450

- *The bottom 54 research output universities* The domain model again performs well but the file and directory models are also good in the educational model.

Table 7
Normality tests (top figure in each cell) and standard deviations (bottom) for the bottom 54 research output universities. Italic cells are significantly non-normal and the bold cells have the smallest standard deviation in each column.

	edu links to all links ratio	com links to all links ratio
File model	0.200+ **0.0286**	0.200+ 0.137
Directory model	0.200+ **0.0286**	*0.001* *0.110*
Domain model	0.095 0.0308	*0.011* ***0.0652***
University model	0.074 0.0343	*0.036* *0.0697*
File range model	0.175 0.0400	*0.016* 0.116
Directory range model	0.200+ 0.0358	*0.000* *0.101*
Domain range model	0.065 0.0330	*0.010* *0.0672*

5.2. The Australia 2001/2002 Data Set

5.2.1. Internal consistency tests: standard deviations and normality tests for ratios of specified outlinks to all outlinks

- *All universities* The domain models perform well, with the domain range version giving slightly more internally consistent results than the standard domain model.

Table 8
Normality tests (top figure in each cell) and standard deviations (bottom) for all Australian universities. Italic cells are significantly non-normal and the bold cells have the smallest standard deviation in each column (n=38).

	edu links to all links ratio	com links to all links ratio
File model	0.200+ 0.0329	0.200+ 0.1092
Directory model	0.200+ 0.0253	0.200+ 0.0730
Domain model	0.200+ 0.0213	0.200+ **0.0518**
University model	*0.135* *0.0221*	0.200+ 0.0557
File range model	0.200+ 0.0276	0.200+ 0.0946
Directory range model	0.200+ 0.0236	0.200+ 0.0673
Domain range model	0.105 **0.0207**	0.200+ 0.0539

5.3. The New Zealand 2002 Data Set

5.3.1. Internal consistency tests: standard deviations and normality tests for ratios of specified outlinks to all outlinks

- *All universities* These results do not exhibit any degree of consistency between the com and edu variations, possibly as a result of the relatively small sizes of web sites but more likely as a result of the very small data set.

Table 9
Normality tests (top figure in each cell) and standard deviations (bottom) for all NZ universities. Italic cells are significantly non-normal and the bold cells have the smallest standard deviation in each column (n=8).

	edu links to all links ratio	com links to all links ratio
File model	0.200+ **0.0246**	0.139 0.154
Directory model	0.200+ 0.0264	0.090 0.0835
Domain model	0.192 0.0284	0.074 0.0662
University model	0.200+ 0.0311	*0.014* *0.0649*
File range model	0.200+ 0.0400	*0.014* *0.134*
Directory range model	0.200+ 0.0302	0.200+ 0.0500
Domain range model	0.200+ 0.0310	0.200+ **0.0472**

6. Discussion

6.1. Are the internal and external consistency tests broadly consistent?

The external consistency tests for the UK gave ambiguous results, with very similar correlations within all columns of all three tables 2-4. The file range model performs consistently well for the whole data set and the top half but the domain model appears to be slightly better for com links. The lack of decisive results is disappointing. It is perhaps a consequence of the necessity to use the weaker rank-based Spearman test rather than Pearson. Nevertheless, since all reported correlations are highly significant it must be concluded that all counting models are effective at producing a pattern from the data to some degree.

The internal consistency models for the UK pointed towards the use of the domain model for the whole set or most productive half, but the edu ratio test cast some doubt upon this for the bottom half. The data for the top 53 institutions is apparently more normal than for the bottom 54, probably due to the smaller numbers of links involved in the latter case. The recommended model from the internal perspective is the standard domain model, especially for the top half of the data set. Since the internal consistency tests are more clear-cut than the external, it is recommended that the document model is used for this data set. The answer to the first research question is in the negative, however, because the two testing approaches give inconsistent results. Some possible explanations for this are suggested.

- The external tests are not powerful enough to discriminate effectively due to their reliance upon ranks.
- The best fitting model for one type of link (e.g. edu links) may be different from another (e.g. com links)
- The ITTD hypotheses may simply be too poor a fit to the data to give reasonable results.

In the Australia data set the domain model again performs well for com links, and this time the domain range model is slightly better for edu links. The use of either is thus

recommended. The New Zealand data set breaks the domain model pattern for internal links but the inconsistency of the results within and between columns is indicative that there are too few universities to make a reliable judgement.

The model that has performed best overall is the standard domain version and so this is recommended as the default for analysing national university systems, principally on the basis that its results are the most internally consistent for UK and Australian TLD target counting. This recommendation does not have the statistical backing of a successful hypothesis test, however, nor the security of validation by external tests and so in no sense represents a robustly validated conclusion.

6.1.1. National reporting and international comparisons

Figures 2 to 4 show the median proportion of site outlinks to the most commonly targeted TLD for each country. In fact this is only one of many relevant sets of statistics that could be displayed, but for reasons of space this is the only type of descriptive statistic that will be reported.

This data can be analysed from two perspectives: each graph can be taken in isolation as a report on the nation concerned; or compared with the others to give an international dimension. The latter is probably instructive even if the focus of a study were to be one particular nation as the figures have more meaning when contrasted with others from a similar source. In this context it is clear that there are trends across the countries but also national differences. It is evident that com edu and org links feature at the top in all cases, in descending order. The host nation is mixed with these, appearing second in two but first in the UK. Links to the com domain form an almost identical proportion in each case. It can also be seen that in addition to the gTLD domains, the UK, Canada and Australia figure in all cases, but New Zealand only in its own graph. All these four countries are industrialised partly or wholly English speaking countries with common historical roots and in a sense, therefore, natural targets. This model is centred only on the source domain and does not take into account the size of the target domain and so New Zealand's small size may account for its non-appearance. Interestingly, an alternative technique (Thelwall & Smith, 2002) that does take into account source and target size found that academic interlinking in the Asia-Pacific region measured on this basis was the greatest between the countries with the smallest web presence, showing that even this more 'balanced' approach has problems. It also shows the need for clarity about what the graphs can and cannot be used to show. For example, New Zealand on an absolute level is not a significant target for UK and Australian universities but may well not be true that individual New Zealand universities are relatively infrequently targeted by the UK and Australian.

Germany and the USA (through the edu domain) feature in all lists. It is known that the USA is the central player on the web, but this is evidence for the importance of Germany, which cannot claim a more significant cultural connection with the three nations than any other country in Europe, for example.

From their linking patterns, New Zealand and Australia appear to be academically focussed on Europe and North America. Japan, for example, is a surprising omission.

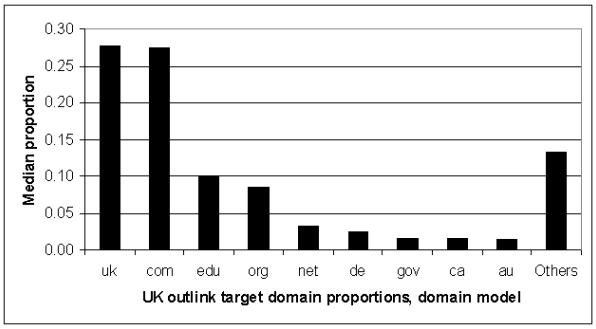

Fig. 1. Median proportions of site outlinks to each of the most highly targeted TLDs for the 109 UK universities.

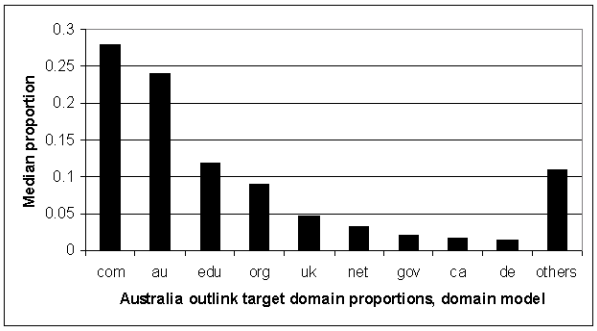
0.3
0.25
0.2
0.15
0.1
0.05
0
Median proportion
com
au
edu
org
uk
net
gov
ca
de
others
Australia outlink target domain proportions, domain model

Fig. 2. Median proportions of site outlinks to each of the most highly targeted TLDs for the 38 Australian universities.

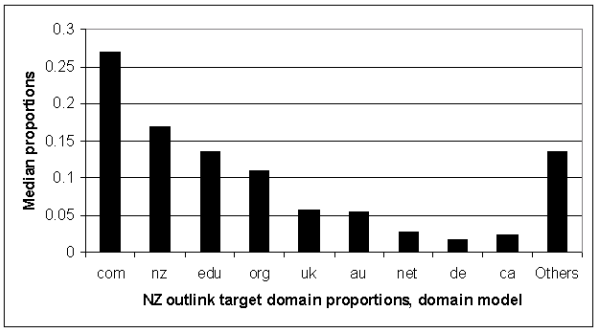

Fig. 3. Median proportions of site outlinks to each of the most highly targeted TLDs for the 8 New Zealand universities.

6.1.2. Anomaly identification

The medians shown in Figure 1 to 3 can be used as national benchmarks and individual institutions compared against them. This could be used as a technique to identify anomalous linking patterns in an investigation of universities. Two outliers are selected from each country to illustrate the potential (domain model).

- NZ/Aukland University of Technology. 9% of links target within NZ, compared to a median of 17% for all NZ universities.
- NZ/Aukland University of Technology. 36% of links target com, national median 27%
- UK/Surrey College of Art and Technology. 88% of links target within the UK, national median 28%.
- UK/Exeter University. 17% of links target de, national median 2%.
- Australia/Bond University. 44% of links target com, national median 28%.
- Australia/University of Ballarat. 55% of links target within Australia, national median 24%.

From these examples it can be seen that some universities have patterns very different from the median. For example, Bond University appears to have a very international commercial orientation whereas Ballarat and Surrey College of Art and Technology seem to be much more nationally focussed. Exeter University is also strange for its German links. These trends really need to be verified by tracking down the source of the links. In the case of Exeter the individual links were scanned and extensive linking to German sites was found throughout the Exeter domain, including links to preprint archives for maths papers and to German physics research groups. A single cause of the particularly high link count was identified, however, "Exeter University German Media Index" (http://www.ex.ac.uk/flc/germtips/media.html), which claimed to host "730 pages of links to German language newspapers, magazines, news-ticker services, TV and radio stations all over the world". Exeter's German Department seems to be particularly well connected on the web, but the particular form of resource that they host frankly foils all of the document models since its targets are likely to be mainly on different web sites. The problem, then, is that the exceptionally high level of (mainly) single TLD multiple site outlinking from a single resource violates the ITTD hypotheses to a great degree. It does not really make sense to report the Exeter TLD proportions as average linking in any meaningful way, even when the domain model is used. As a result Exeter must be treated as an anomaly, the type of which presents a further challenge for future document model construction. National medians should not be greatly affected by anomalies such as this one, however, as long as they are relatively rare.

6.2. Reporting the Outlink Results: Do the Models Make a Difference?

An important question is whether the model of web linking used makes a difference to the results. Figure 4 shows the results of all models for the UK and it can be seen that the difference in proportions between models is actually very big. This can be seen clearly both in the absolute size of the UK columns and in the relative heights of these and the com columns.

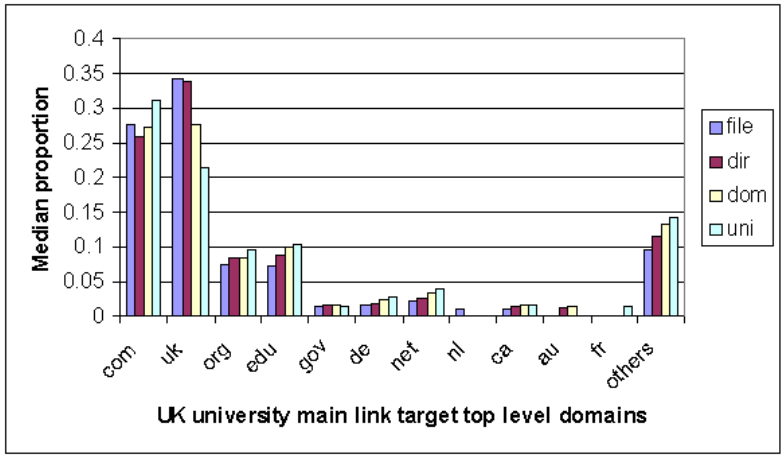

Fig. 4. Median proportions of site outlinks to each of the most highly targeted TLDs for the four different domain models in the UK. Missing columns indicate TLDs not falling into the top 9 for the model.

7. Conclusion

The task of objectively assessing the average targets of links from universities is clearly a problematic one. The domain model performed the best overall in the tests to assess which link counting methodology was internally the most consistent. The example of the UK showed that the choice of model could make a large difference to the actual results (Figure 4) but the external validity test did not endorse this choice for the UK and not one of the results was clear-cut.

In terms of the underlying processes, the success of the domain model was not emphatic enough to be able to claim in any sense that it is the 'correct' model for objectively reporting on the patterns of link target selection by universities, even allowing for a few aberrations. A fairer characterization would be that it is the least badly fitting of the models, but it value lies in being significantly better than the default file model. In reality, web sites will simultaneously contain 'documents' in a variety of forms: in files, directories, domains and across-domains. The key issue for this study of identifying when multiple links 'should' be counted as one is likely to be clear cut in some cases (e.g. two copies of the same file in different locations) and ambiguous in others (e.g. a list of links to twenty papers by the same author on another web site). Fundamentally, if humans were to be used then value judgements would have to be made and, therefore, for an algorithm oversimplifications are a given. The sheer size of the data sets used means that there is no contest about which approach must be used and so the issue is to develop increasingly effective heuristics. Those tested here are relatively simple, but of some use. A more sophisticated approach might be to fit web files, where possible, into a recognised flexible genre type (e.g. Rehm, 2002) so that one single level of aggregation would not have to dominate an entire site. Automatic genre identification is still in its infancy, however, and has yet to prove that it is a practical possibility. Future work could also involve applying the techniques to groups of large sites outside the academic domain.

In summary, the techniques introduced have not yet proved their worth, although it is claimed that the results should be treated as best estimates. This is unfortunately the most that can be said about many web-related statistics (e.g. Lawrence & Giles, 1999) because of the

difficulty, and in some cases impossibility, of producing statistically valid estimates due to the heterogeneous and uncontrolled nature of the web.

8. Acknowledgement

The author would like to thank Blaise Cronin for many helpful comments on an earlier draft of this paper.

9. References

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

Björneborn, L. (2001). Necessary data filtering and editing in webometric link structure analysis. Royal School of Library and Information Science.

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

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.

Brunn, S. D. & Dodge, M. (2001). Mapping the "worlds" of the world wide web: (Re)Structuring global commerce through hyperlinks, *American Behavioral Scientist,* 44(10), 1717-1739

Cronin, B. & Shaw, D. (2002, in press). Banking (on) Different Forms of Symbolic Capital, *Journal of the American Society for Information Science & Technology*.

Cronin, B., Snyder, H.W., Rosenbaum, H., Martinson, A. & Callahan, E. (1998). Invoked on the web. *Journal of the American Society for Information Science*, 49(14), 1319-1328.

Crowston, K., Kwasnik, B. H., Nilan, M. & Roussinov, D. (2000). Identifying document genre to improve Web search effectiveness. *Proceedings of the 63rd Annual Meeting of the American Society for Information Science*. 37, 124.

Crowston, K. & Williams, M. (2000). Reproduced and emergent genres of communication in the world wide web, *The Information Society*, 16(3), 201-15.

Dearing, R. (1997). Report of the national committee for enquiry into higher education. http://www.leeds.ac.uk/educol/ncihe/

Education Guardian (2001). About the tables, http://education.guardian.co.uk, Accessed 17 December, 2001.

Egghe, L. (2000). New informetric aspects of the Internet: some reflections - many problems. *Journal of Information Science*, 26(5), 329-335.

Flake, G. W., Lawrence, S., Giles, C. L. & Coetzee, F. M. (2002). Self-organization and identification of web communities, *IEEE Computer*, 35, 66-71.

Garrido, M. & Halavais, A. (2002, to appear). Mapping Networks of Support for the Zapatista Movement: Applying Social Network Analysis to Study Contemporary Social Movements. *In*: M. McCaughey & M. Ayers (Eds). *Cyberactivism: Critical Practices and Theories of Online Activism*. London: Routledge.

Glänzel, W. (2001). National characteristics in international scientific co-authorship relations. *Scientometrics*, 51(1), 69-115.

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 & Management*, 37(5), 661-676.

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

Kling, R. & McKim, G. (2000). Not Just a Matter of Time: Field Differences in the Shaping of Electronic Media in Supporting Scientific Communication. *Journal of the American Society for Information Science,* 51(14), 1306-1320.

Koku, E., Nazer, N. & Wellman, B. (2001). Netting scholars: Online and offline, *American Behavioral Scientist,* 44(10), 1752-1774.

Lawrence, S. L. (2001). Online or invisible? *Nature*, 411(6837) 521.

Lawrence, S. & Giles, C. L. (1999). Accessibility of information on the web. *Nature*, 400, 107-109.

Nicholas, D., Huntington P., Lievesley, N. and Withey, R. (1999). Cracking the code: Web log analysis. *Online & CD-ROM Review* 23(5), 263-269.

Nilan, M. S., Pomerantz, J., & Paling, S. (2001). Genres from the bottom up: What has the Web brought us? *Proceedings of the 64th Annual Meeting of the American Society for Information Science and Technology*. 38, 330-339.

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 & Technology,* 53(7), 592-601.

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.

Rehm, G. (2002). Towards automatic web genre identification. In: *Proceedings of the 35th Hawaii International Conference on System Sciences*, (2002). (http://www.computer.org, electronic restricted access document).

Smith, A. & Thelwall, M. (2002). Web Impact Factors for Australasian Universities, *Scientometrics*, 54(1-2), 363-380.

Spink, A., Wolfram, D., Jansen, B. J., & Saracevic, T. (2001). Searching the web: The public and their queries. *Journal of the American Society for Information Science*, 53(2), 226-234.

Stock, G. (2002). Googlewhacking: The Search for The One True Googlewhack, Available: http://www.googlewhack.com/, accessed March 27, 2002.

Tang, R. and Thelwall, M. (2002, in press). Exploring the pattern of links between Chinese university Web sites, Proceedings of the ASIST Annual Meeting Volume 39 (ASIST 2002).

Thelwall, M. (2001a) Web Log File Analysis: Backlinks and Queries, *ASLIB Proceedings, 53(6) 217-223*.

Thelwall, M. (2001b). Extracting macroscopic information from web links. *Journal of the American Society for Information Science and Technology*, 52 (13), 1157-1168.

Thelwall, M. (2001c). A publicly accessible database of UK university website links and a discussion of the need for human intervention in web crawling, University of Wolverhampton. Available: http://www.scit.wlv.ac.uk/~cm1993/papers/a_publicly_accessible_database.pdf.

Thelwall, M. (2001d). Exploring the link structure of the Web with network diagrams, *Journal of Information Science* 27(6) 393-402.

Thelwall, M. (2001e). A web crawler design for data mining, *Journal of Information Science*, 27(5), 319-325.

Thelwall, M. (2002a, in press). The top 100 linked pages on UK university Web sites: high backlink counts are not usually directly associated with quality scholarly content, *Journal of Information Science*.

Thelwall, M. (2002b, in press) A bibliometric investigation into Google's PageRank algorithm applied to national systems of university websites, *Journal of Documentation*, 58(6).

Thelwall, M. (2002c) 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. (2002e). A comparison of sources of Links for academic Web Impact Factor calculations. Journal of Documentation, 58, 60-72.

Thelwall, M. (2002f, in press). A Research and Institutional Size Based Model for National University Web Site Interlinking, *Journal of Documentation*.

Thelwall, M. (2002g). New Linking Motivations? General Navigational, Ownership, Social and Gratuitous Links, *University of Wolverhampton*.

Thelwall, M., Binns, R., Harries, G., Page-Kennedy, T., Price E. and 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(4)*.

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

Thelwall, M. & Wilkinson, D. (2003, in press). Three Target Document Range Metrics for University Web Sites, *Journal of the American Society for Information Science and Technology*.

van Raan, A. F. J. (2001). Bibliometrics and Internet: Some Observations and Expectations, Scientometrics, 50(1), 59-63.

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

Wilson, T. (2002). Web citation. JESSE listserv discussion. Available: http://listserv.utk.edu/cgi-bin/wa?A1=ind0205&L=jesse