Graph Structure in Three National Academic Webs: Power Laws with Anomalies

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

The graph structures of three national university publicly indexable webs from Australia, New Zealand and the UK were analysed. Strong scale-free regularities for page indegrees, outdegrees and connected component sizes were in evidence, resulting in power laws similar to those previously identified for individual university web sites and for the AltaVista-indexed web. Anomalies were also discovered in most distributions and were tracked down to root causes. As a result, resource driven web sites and automatically generated pages were identified as representing a significant break from the assumptions of previous power law models. It follows that attempts to track average web linking behaviour would benefit from using techniques to minimise or eliminate the impact of such anomalies.

1 Introduction

The Web can be modelled as a mathematical graph by considering its pages to be nodes connected by arcs corresponding to hyperlinks. The study of the structure of this graph is useful because of the importance of hyperlinks for search engine web crawlers and in information science web link research (e.g. Björneborn, 2001; Thelwall, 2001a). For example Web topologies are studied in the Borgman & Furner (2002) review of bibliometrics. The topology of the Web is also important background information for the many researchers in different disciplines that are interested in visualizing or exploiting link structures (e.g. Brunn & Dodge, 2002; Jung *et al.*, 2002; Garrido & Halavais, 2002; Thelwall & Smith, 2002). There have been several previous attempts to analyse the web from a purely graph theoretical point of view, perhaps triggered by the theory of ‘small world’ phenomena in networks (Watts & Strogatz, 1998). Albert *et al.* (1999) analysed the structure of the Notre Dame University web site, conjecturing that in the whole web, the average hyperlink based distance between any two pages at random was 19. However, Broder *et al.* (2000) subsequently showed that the web was in fact disconnected so that many pages could not be reached. Huberman & Adamic (1999) found power laws in the distributions of total pages in sites covered by Alexa and Infoseek crawls. A power law occurs above when the frequencies *n* of some variable x, such as total pages on sites, are proportional to $1/x^n$. There have been several studies on the distribution of links. Rousseau (1997) found a power law for links counts from AltaVista to a topic-based set of pages. Barabási *et al.* (1999) showed that power laws apply to vertex connectivities in many large networks, including the web. It was conjectured that this was a result of a combination of continual network expansion and the preferential

[1] *This is a preprint of an article accepted for publication in the Journal of the American Society for Information Science and Technology*

attachment of new links to pairs of pages that were already well connected. Preferential attachment has been known in information science at least since Price (1976), with descriptions such as 'success-breeds-success'. Adamic & Huberman (2000) claimed that the explanation, but not the model, was a poor fit for the data because older sites did not receive significantly more links than newer ones. Barabási *et al*. (2000) countered that such a trend could be present if more averaging were applied. It is clear, however, that although the model is a good fit, the temporal explanation for this fit is less so. A solution was provided by the revised model of Pennock *et al*. (2002), which combines a power law with random linking. This makes it easier to incorporate the possibility for new sites to gain high link counts. This model also explains the partial non-linearity of many power graphs and it was shown that the type of domain selected has a great impact on the relative effect of the competing tendencies for random links and links going preferentially to already well connected pages. The inbound links of university home pages, for example, show more tendency for random links than preferential attachment. This may reflect links characterised by awarding 'credit' to the institution rather than recognising page content usefulness (Thelwall, 2002b).

The most detailed report so far of the overall link structure of the web has been that of Broder *et al*. (2000) based upon data from two AltaVista crawls, described in detail below. The outcome of this research was an extensive description of many aspects of the connectivity of the area covered, including a pictorial representation of its overall structure. It is now proposed to apply these techniques to systems of university web sites with the twofold aims of getting specific information about this important sector of the web, and to make the results more concrete in terms of identifying causes for unusual phenomena observed through visiting the pages involved. Academic web links have been extensively discussed and studied in information science and so are a natural choice (Rousseau, 1997; Ingwersen, 1998; Leydesdorff & Curran, 2000; Cronin, 2001; Borgman & Furner, 2002). It is particularly important from a web metrics perspective to be able to identify regular and anomalous behaviour in web spaces so that steps can be taken to minimise or eliminate their impact on results (Thelwall, 2001a). The web systems chosen are three large mature academic web spaces: from Australia, New Zealand and the UK.

2 The AltaVista Study

Broder *et al*. (2000) used two complete crawls from AltaVista loaded into a purpose-built program on a large computer to investigate the connectivity of the web. Their principal discovery was that their crawl data could be split into five parts: IN; OUT; Strongly Connected Component (SCC); TENDRILS; and DISCONNECTED. The first four had roughly equal sizes. A strongly connected component in a (directed) web graph is a collection of pages from which a crawl following only links in pages could start anywhere in the set and reach every other page in the set. The SCC is the largest such component. OUT is the set of pages outside SCC that can be reached from all SCC pages but do not connect back to any page in the SCC. IN is the set of pages that connect to the SCC but are not connected to it so a crawl starting in IN would contain all of the SCC and OUT, plus some additional pages from IN. TENDRILS is a separate collection of pages that are either linked to by a page of IN or link to a page of OUT but are not in IN, OUT or SCC. Finally, DISCONNECTED is the set of pages that are not linked in any way to the other four components. See Baeza-Yates, R. & Castillo (2001) for a more detailed component breakdown that is possible using the same techniques.

3 Crawl Coverage: Commercial Search Engines and Academic Web Link Research

There are many methodological issues arising from the AltaVista study when transferred to an information science context. One puzzle is how the pages outside SCC and OUT were found, since they were not linked to via pages in the core of the web. A crawl starting at a single point in SCC would not find any of these pages. Some of them come from URLs submitted to AltaVista by website owners (Broder *et al.* 1999). Others probably were connected to by indexed pages in the past but are now no longer connected but are still recorded in the database and visited by the crawler. This means however, that researchers without access to a major search engine database will only be able to study SCC and OUT systematically, an obvious limitation.

A crawler cannot guarantee to identify all outlinks in a web page that it has fetched. In particular links created by JavaScript, server side image maps and embedded applications could be expected to be inaccessible (Thelwall, 2002a). This means that the AltaVista results only apply to the indexable links: all the components except DISCONNECTED could be larger in reality from this cause. It also provides an additional reason for AltaVista being able to index pages outside SCC and OUT: they could have previously been connected through standard links with these subsequently having been converted to non-indexable links.

A related problem is that some links may be ignored as a policy decision. For example particular links that appear to be a database query (including an '?' in the URL) could be systematically ignored as could frameset pages because of the difficulty in using them to suggest meaningful pages to search engine users. Search engines also have secretive policies for banning spam sites, and policies can change with time (Sullivan, 2002).

The AltaVista data set was based upon the crawl before removing "duplicates and near duplicates". Duplicates web pages exist in many places on the web, from individual pages to entire sites. Sets of computer documentation for example, are commonly mirrored frequently across the globe. Since these appear to be commonly very well interlinked, it is possible that their deletion would reduce the size of SCC or OUT.

The AltaVista data set only included HTML pages and not, for example, resources such as PDF files and images that are linked to by other HTML files (Sullivan, 2001). This is one logical choice but another would be to include all resources linked to by web pages, including those that are of a type that does not contain links. Google, for example, indexes PDF files and many search engines now incorporate image search capability.

4 Method - Analysis Environment and Data Sets

The AltaVista results will be compared with some mature academic web spaces. The structure of Australian (crawled October 2001 to January 2002), New Zealand (January 2002 to February 2002) and UK (July 2001) universities' publicly indexable web sites were obtained from a previously used freely available source (http://cybermetrics.wlv.ac.uk/database). The publicly indexable collection consists of all pages reachable from a crawl starting at each university's home page, or other similar start point. The crawler used is described in detail in Thelwall (2001b). 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. A program was then used to identify, through domain names, all links targeted at recognised same country university web sites, removing the rest. From recognised links the URLs were converted into purely numerical format, producing two files; one the link structure in numbers and the other a key for identifying the actual URL represented by each number. A program was then written, *Graph Structure Analysis Environment*, to load and hold the entire graph simultaneously in main memory and to execute a range of graph analyses on it, including Breadth First Searches and connectivity tests.

The data source differs in several important ways from that of AltaVista.

- The crawls of each university site were independent of each other so that a link from one university to another would not have been used to instruct the crawler to visit the target page.
- Non-HTML linked pages were included, e.g. those in Portable Document Format.
- Duplicate checking was performed at crawl time so that identical pages with different URLs were rejected. Any links to duplicate pages would not be redirected to the alternative location, however.
- No historical information was used from previous crawls about the URLs visited: a single starting URL would have been used in each case, normally the home page.
- A different list of banned pages and areas would inevitably have been used. For example mirror sites were excluded, when identified.
- Fine details of the crawling could be expected to be different, although it is not possible to be specific about this since AltaVista does not publish its algorithm. For example, it is suspected that AltaVista may ignore frame-based pages in some cases.

5 Findings

5.1 Indegrees and Outdegrees

Table 1 gives some basic descriptive statistics extracted from the three systems. These are discussed in more detail below.

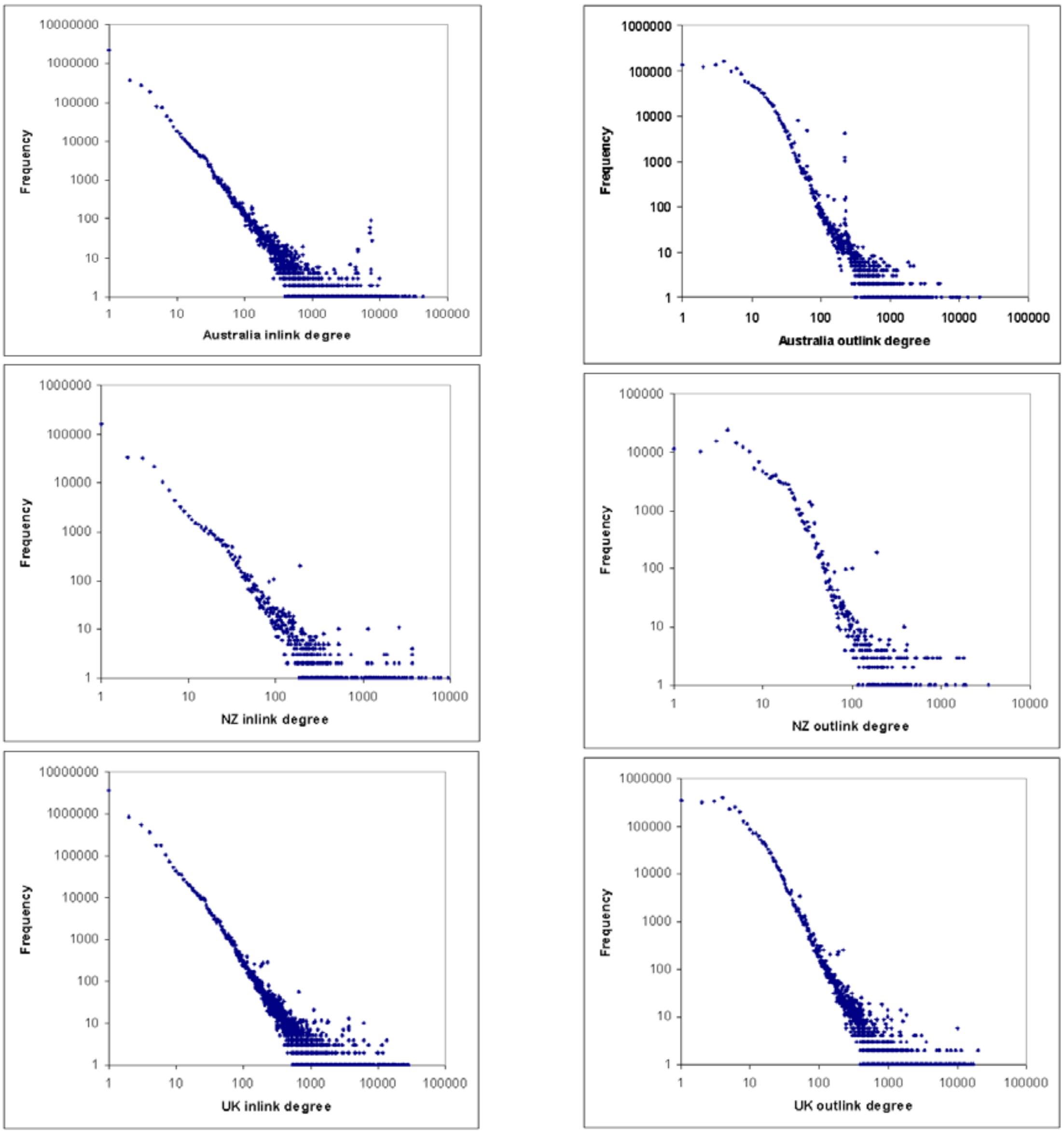
Table 1. Basic statistics for each national system. All figures include only inlinks and outlinks to and from recognised university sites in the same country. *OUT inlinks include all links from SCC and SCC outlinks include all links to OUT.

Component	Australia	New Zealand	UK
OUT (%)	2,548,276 73%	213,078 70%	4,557,998 70%
SCC (%)	963,231 27%	92,102 30%	1,995,602 30%
Total pages	3,511,507	305,180	6,553,600
Total links	18,031,706	1,874,141	31,250,705
Maximum number of links to a page (inlinks)	42,903	9,599	27,897
Median number of links to a page (all/SCC/OUT*)	1/2/1	1/3/1	1/3/1
Maximum number of links from a page (outlinks)	20,000	3,378	19,999
Median number of links from a page (all/SCC*/OUT)	0/8/0	1/7/0	0/6/0
Number of different indegrees	1,465	523	1,686
Number of different outdegrees	947	298	1,461
Longest shortest directed path A->B	362	1,445	1,022

Figures 1-6 show plots of indegrees and outdegrees, showing that these adhere to approximate power laws. The indegree of a page is the number of links to it and its outdegree is the number of links from it, i.e. contained within its HTML. It should be remarked, however, that pages with zero outdegrees had to be excluded before both plotting the three outlink graphs and fitting the outdegree lines. A slightly hooked shape is evident, albeit to differing degrees, at the top left of each graph. This can be explained in terms of the Pennock, Flake, Lawrence, Glover & Giles (2002) model (PFLGG model). This is a theoretical mathematical model of the growth of edges in a network, building on the work of Barabási & Albert (1999). The underlying assumption is that links will be added to the web in a non-random fashion preferentially attaching to pages that are already well connected. The pure version of this would result in a straight-line graph and a perfect power law. With the PFLGG model there is another competing factor at work, links being added to pages unaffected by their existing connectivity. The hook shape in the graph indicates that this latter tendency is present, but is still dominated by preferential attachment. Interestingly, however, the outlink graphs are all significantly more hooked than those for inlinks, although the PFLGG is a symmetrical model. A logical explanation for this is that the choice of page to link to is significantly more affected by its connectivity than the choice of page to link from. In other words, even badly connected pages will tend to host a few links – more than the preferential attachment model would predict. Broder *et al.* (2000) commented that when applying the Zipf rank-based model to the data the noticeable flare out at the bottom of each graph disappears and the graph becomes a straight line, but this does not really help to explain the phenomenon. The flare is actually a natural breakdown of the graphical representation of the model and is understandable if the power law is seen as a probabilistic prediction. The long line on the frequency = 1 axis of each graph is actually a sparsely separated set of points, although this is not clearly visible from the graphs. In this context the absence of dots is as significant as their presence. So, for example, if there is only one point between 11,000 UK inlinks and 12,000 UK inlinks then this would be entirely consistent with each inlink count between these two figures having a probability of 0.001 of occurring.

There are a number of anomalies in each graph. The main New Zealand anomalies came from a set of highly interlinking software documentation. From Australia, the biggest came from a single source: the University of New South Wales's online course handbook. Most of the thousands of course pages had a standard navigation bar. This accounts for the very high similar outdegrees and the set of "AtoZ guide" link targets common to all of the pages account for the high inlink counts. The links were operated through a dynamic HTML menu so that they would not all be visible at the same time and clutter up the screen, but they were nevertheless embedded in the text as plain HTML (using 'layers') and so could be extracted by the crawler. This allows the pages to have many more outlinks than could be expected from a normal navigation bar. In the UK the four huge inlink counts all came from the standard navigation bar on the City University Atheneum Project author index. There is a separate HTML file for each author associated with the Atheneum magazine that the project has indexed. The two large outlink counts are both associated with a single bioinformatics database results indexes. The common theme is that all of the biggest anomalies are produced by internal links within data-driven sites. These can easily be seen to be violating the hypotheses of the PFLGG model.

Comparing indegrees and outdegrees between SCC and OUT (not illustrated), some clear differences emerge. Firstly, all SCC pages have outdegree of at least 1, by definition, whereas the median outdegree of OUT pages is 0. The median for SCC of between 6 and 8 shows that there is significantly more outlinking from within SCC pages. The same is true for inlinking, but to a lesser degree. In fact inlinking and outlinking from both SCC and OUT display power laws, so although there are generally more links within SCC, OUT also has a spectrum of the more highly connected pages. As a final point, a median of 2 or 3 inlinks for SCC shows that this area is actually very sparsely interconnected. The average SCC page can only be reached from 2 (Australia) or 3 (UK, New Zealand) other SCC pages.



FIGs 1-6. These six graphs are logarithmic displays of inlink and outlink counts for Australia, New Zealand and the UK, showing the number of pages that have each indegree and outdegree. The linear shapes indicate the workings of approximate power laws. Note that all pages have an inlink count of at least one, but pages with an outlink count of 0 cannot be displayed and are omitted from the graphs.

5.2 Weakly and strongly connected components

5.2.1 Weakly connected components

For each data set the component OUT was extracted and its internal connectivity assessed by a partition into weakly connected components. A weakly connected component is a set of pages such that all pages can be reached by all other pages by traversing a set of links in any direction or combination of directions. Graphs plotted of the component sizes show a very similar power law pattern, with the exception of the excess of components of size 1. Figure 7 shows the graph for the UK OUT component, the others having a very similar shape. In all countries just under half of the nodes in OUT do not link to any other node in OUT and are not linked to by any OUT node. These are relatively isolated web pages that are linked to by at least one SCC page but not by any other pages.

The largest undirected components were explored and the causes traced back. Several were huge databases of academic resources. There were several sets of computer documentation in this group. Some further illustrative examples are given below.

- 140,932 pages from a costal imaging pictures database in Australia. Most of these are stored as pure pictures, jpeg files.
- 88,437 informational pages from a DNA database site in the UK.
- 23,030 pages forming the archives of the "Information Bulletin on Variable Stars (IBVS)".

There were also some strange undirected components.

- 262,798 pages from a mixture of different and unrelated database sites in the UK.
- 19,124 pages mainly from a picture gallery and an online statistical software manual.
- 2,869 automatically generated pages consisting of feedback forms. Each page URL contains a unique number and a link to an identical URL but with the number incremented. This link serves to generate a new page. All the pages are identical except for the link URL to the next page, which has (normally) only one digit different. This theoretically infinite process only terminated when the remote web server stopped responding.
- 57,182 pages from a single directory holding pages containing incorrectly formatted relative links. The result of the incorrect formatting was to add the current directory to the current path instead of substituting it. This operated recursively with the crawler to generate enormous URLs. For example part of the URL of one page was "source/io/source/", where the source directory had been appended to the end of the io directory instead of replacing it. The incorrect URLs were intelligently interpreted by the web server and appropriate pages delivered so that the errors could be compounded. Each page contained variable components and so no duplicate elimination was triggered.

The graph shows a breakdown of the power law for component size n = 1, perhaps due to the inclusion of non-HTML resources. The huge number of components of size 1 is interesting (2,220,070, containing 49% of pages for the UK) because these must all be linked to by pages from the SCC, indicating that the SCC must be surrounded by a fuzz of individual pages that do not link to any other national university pages. Many of these will probably be non-HTML resources that cannot contain crawled links.

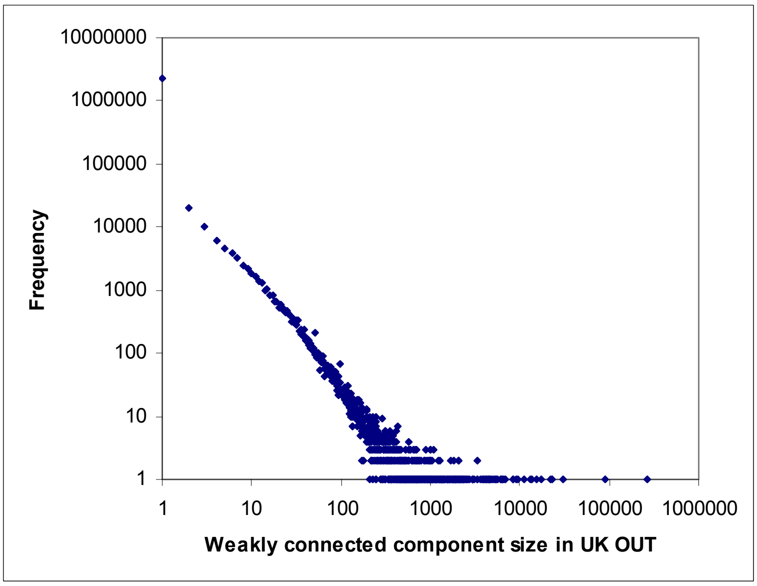


FIG. 7. Sizes of the weakly connected components in UK OUT approximate a power law, except for those of size 1.

5.2.2 Strongly connected components

The strongly connected components of each country formed a similar pattern to each other, with a power law and an extreme point on each axis. There are a huge number of components of size one, web pages that are not in any loops of links between pages within the same national university system. This could be partly explained by the inclusion of non-linking resources in the data set, such as images, all of which would form components of size one. The other extreme point was formed in each case by the SCC for each country. In the UK and Australia it was two orders of magnitude bigger than the second largest component but in New Zealand only 23 times as big.

The largest components other than the SCC were investigated for each country and were typically found to be self-contained sets of computer documentation that had extensive internal navigation links. The exceptions were: a Leeds University undergraduate module information catalogue; a web site for flags of the world; and the aforementioned automatically generated feedback forms. Although the largest component in each country is heterogeneous and cross-site, its nearest neighbours are very homogeneous and single site. From this it seems that seeking large strongly connected components is unlikely to give very revealing information about the web, and that the weakly connected approach is more robust to the vagaries of linking, although it can only be applied after the SCC has been eliminated.

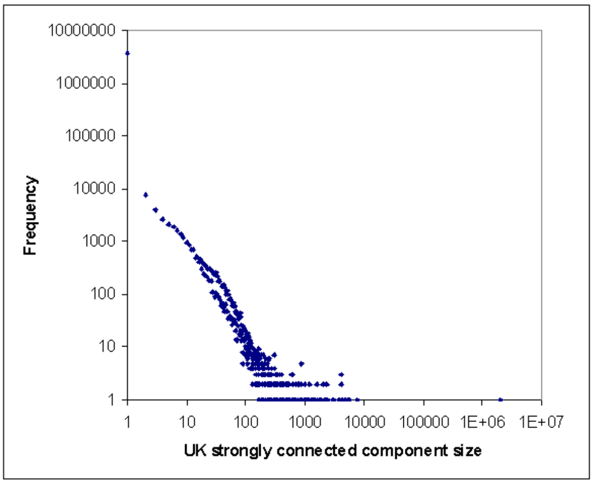


FIG. 8. Sizes of the strongly connected components in the UK approximate a power law, except for an extreme point on each axis.

5.3 Longest shortest paths in graphs

For any pair of pages it is possible to ask what is the least number of links that must be traversed to get from the first one to the second. This is called the shortest path between them. For a whole set of pages it is then of topological interest to know what is the longest of all these shortest paths. Each of the national university systems was checked for the longest shortest path. This was achieved by taking a university home page for the start of a crawl and then identifying the last page(s) reached in a breadth-first search, then crawling backwards from these pages to identify the new last page. This does not guarantee to get the longest shortest path since a different initial starting point could possibly yield a different one, but it gives a lower bound for the longest shortest path. The results were as follows.

- In Australia the longest shortest paths found were of 362 links from http://itee.uq.edu.au/~comp1001/slides/267.html to http://medieval.unimelb.edu.au/ductus/demo/files/ductus/frames/cappelli/57.control.html and http://medieval.unimelb.edu.au/ductus/demo/files/ductus/frames/cappelli/57.paper.html
- In New Zealand the longest shortest path found was of 1,445 links from http://cgr.otago.ac.nz/slides/prochip/tsld023.htm to http://webview.massey.ac.nz/scripts/idparser.dll/000000100000010 1403/comment.html. The tail of this coincides with part of the large undirected component described above.
- In the UK the longest shortest paths found were of 1,022 links from http://www.hud.ac.uk/schools/design_technology/textiles/show98/baxen006.htm to http://papaya1.ncl.ac.uk/wb_5_8_2019.html and http://papaya1.ncl.ac.uk/wb_9_1_1984.html.

From these results it can be seen that very long paths do exist in the data set, with the end pages being buried deeply in obscure places.

6 Summary

Power laws are clearly very evident in many aspects of the topology of national university webs, as in other areas previously examined. In particular there is evidence for a rich get richer model of new links being disproportionately added to pages that already host many and targeted at others which are already popular targets. However, there is evidence for a small degree of a countervailing tendency to link at random, consistent with the PFLGG model. Despite the strong regularities found, anomalies were also present from three causes.

- Automatically generated pages served to the crawler, and those produced by automatically fixed link errors.
- The inclusion of non-HTML web ‘pages’, in particular because these cannot host links, or the crawler did not extract links from them.
- Large resource-driven web sites.

It should be emphasised that the dominant regularities found are in spite of all inlinks and outlinks not associated with recognised national universities being discarded, indicating the robustness of the phenomena identified to a restriction of the domain of consideration. Ignoring the second factor and just considering HTML pages it is concluded that the academic web contains (1) resource-driven subsites and (2) automatically generated pages that are structurally inconsistent with the remainder of the web. In a sense, then, these areas are alien to the web. Of course the academic web is a subset of the whole web but the relatively higher incidence of this problem is evident from a comparison with the graphs of Broder et al. (2000). The implication of this for studies of academic web links is that steps should be taken to either segregate out such areas or employ techniques that are not sensitive to large anomalies in order to get the most meaningful result about the “normal” web. This can be for topological investigations or, with more direct applications, for link metric research. Taking this argument one step further, there is no reason to believe that anomalies come only in large sizes. In reality it is almost certainly common practice for scholars to post a small collection of interlinked pages. This would be an alternative explanation to that of the PFLGG model for the hook shape in graphs: the relative rarity of individuals publishing single web pages rather than small collections. This is more realistic than the PFLGG model, with nodes and arcs being added individually rather than in groups, but does not yield a simple model.

7 References

Adamic, L. A. & Huberman, B. A. (2000). Power-Law distribution of the World-Wide Web, *Science*, 287, 2115a

Albert, R., Jeong, H. & Barabási, A. L. (1999). Diameter of the World-Wide Web, *Nature*, 401, 130-131.

Baeza-Yates, R. & Castillo, C. (2001). Relating Web characteristics with link based Web page raking, In: Proceedings of SPIRE 2001, IEEE CS Press, Laguna San Rafael, Chile, pp. 21-32.

Barabási, A. L. & Albert, R. (1999). The emergence of scaling in random networks, *Science*, 286, 509-512.

Barabási, A. L. Albert, R., Jeong, H. & Bianconi, G. (2000). Response to Adamic and Huberman, *Science*, 287, 2115a

Björneborn, L. (2001). Small-world linkage and co-linkage. In: *Proceedings of the 12th ACM Conference on Hypertext and Hypermedia* (pp. 133-134). New York: ACM Press.

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

Broder, A. Kumar, R, Maghoul, F., Raghavan, P., Rajagopalan, S., Stata, R., Tomkins, A. & Wiener, J. (2000). Graph Structure in the Web, *Journal of Computer Networks*, 33(1-6),309-320.

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

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.

Huberman, B. A. & Adamic, L. A. (1999). Growth dynamics of the World-Wide Web, *Nature*, 401, 131.

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

Jung, S., Kim, S. & Kahng, B. (2002). Geometric fractal growth model for scale-free networks, Physical Review E, 65(5), No. 056101. Available: http://phya.snu.ac.kr/~kahng/PRE56101.pdf

Leydesdorff, L. & Curran, M. (2000). Mapping university-industry-government relations on the Internet: the construction of indicators for a knowledge-based economy, Cybermetrics, 4. Available: http://www.cindoc.csic.es/cybermetrics/articles/v4i1p2.html

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.

Price, D. J. de Solla (1976). A general theory of bibliometric and other cumulative advantage processes. *Journal of the American Society for Information Science*, 27, 292-306.

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

Sullivan, D. (2001). SearchDay October 31, 2001 - Number 128. Available: http://searchenginewatch.com/searchday/01/sd1031-google-files.html, accessed 7 August 2002.

Sullivan, D. (2002). Google Adds More "Fresh" Pages, Changes Robots.txt & 403 Errors, Gains iWon. Available: http://searchenginewatch.com/sereport/02/08-google.html, accessed 7 August 2002.

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. (2002a). Methodologies for Crawler Based Web Surveys, *Internet Research: Electronic Networking and Applications*, 12(2), 124-138.

Thelwall, M. (2002b). 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), 485-493.

Thelwall, M. & Smith, A. (2002). A study of the interlinking between Asia-Pacific University Web sites, *Scientometrics* 55(3), 335-348.

Watts, D. J. & Strogatz, S. H. (1998). Collective dynamics of ‘small-world’ networks, *Nature,* 393, 440-442.