

# An initial exploration of the link relationship between UK university websites

**Mike Thelwall<sup>1</sup>**

*School of Computing and Information Technology, University of Wolverhampton,  
Wulfruna Street, Wolverhampton WV1 1SB*

*m.thelwall@wlv.ac.uk*

Aggregates of links are of interest to information scientists in the same way as citation counts are: as potential sources of data from which new knowledge can be mined. The recent discovery of a correlation between a web link count measure and the research quality of British universities is built upon by applying a range of multivariate statistical techniques to counts of links between pairs of universities. This represents an initial attempt at developing an understanding of this phenomenon. Plausible results are extracted, including the high degree of similarity between the Scottish universities and limited evidence of a dichotomy between new and traditional universities. Outliers in the data were also identified by the techniques, some of which were verified by being tracked down to identifiable web phenomena. This is an important outcome because successful anomaly identification is a precondition to more effective analysis of this kind of data. The identification of groupings is encouraging evidence that web links between universities can be mined for significant results, although it is clear that more methodological development is needed if any but the simplest patterns are to be extracted. Finally, based upon the types of patterns extracted it is argued that none of the methods used are capable of fully analysing link structures on their own.

## Introduction

Web links are potential indicators and generators of trust (Davenport and Cronin, 2000), with a page that is the *target* of many being more likely to contain useful information than one that is not. The analogy with citations is clear and has been remarked upon many times, as has the fact that web links are less reliable than citations due to the lack of quality control over the vast majority of the Web (Egghe, 2000; Bar-Ilan, 2001; Björneborn and Ingwersen, 2001; Thelwall, 2001a). A correlation has been recently discovered, however, between a link count measure and research ratings (Thelwall, 2001a) for British universities. This result shows that meaningful information can be extracted from large-scale comparisons of web links between academic websites, although the results are far from reliable on the level of individual universities. In fact these results were obtained despite the finding that relatively few links were targeted at online academic papers, making the interpretation of link counts highly problematic and necessitating further exploratory research. Given the positive results, there is an exciting opportunity to develop new tools to

---

<sup>1</sup> Thelwall, M. (2002). An initial exploration of the link relationship between UK university web sites, *ASLIB Proceedings*, 54(2), 118-126.

analyse the same kind of data in order to extract more information and to gain a deeper insight into the underlying processes involved. This paper represents a first attempt to develop an understanding of the individual academic website interlinking phenomenon.

There have been several studies of web links in electronic journals and other areas of the Web, many through direct analogy with citations. Studies of e-journals have, so far, not found it possible to get meaningful results from attempts to extend the journal impact factor metric to a backlink-based equivalent (Smith, 1999; Harter and Ford, 2000). It was, however, a generalisation of this, Ingwersen's Web Impact Factor (WIF) that has been shown to correlate with university research ratings when applied to their websites (Thelwall, 2001a; Thelwall, 2002). This example shows that web link aggregation is a technique that has the potential to reveal meaningful underlying trends, in spite of the following serious concerns over many aspects of the validity of such data.

- The search engines often used to collect the data have been found to be unreliable (Bar-Ilan, 1999; Rousseau, 1999; Snyder and Rosenbaum, 1999; Thelwall, 2000), although recent improvements have been identified in AltaVista (Thelwall, 2001b; Thelwall, 2001a).
- Web pages are not subject to quality control in the way that journal articles are.
- Web pages can have multiple URLs. Entire collections of web pages can also be multiply sited in a common process known as mirroring.
- The authors of pages can be difficult to identify, and authorship may, as in the case of many traditional journal articles (Cronin, 2001), be a complex concept.
- Web pages can appear and disappear instantly, even in their thousands, as a result of a single decision (Thelwall, 2001c).
- Counts of web pages are affected by stylistic considerations: an author may publish an online book, for example, as a single huge web page or thousands of small pages.

One previous study has applied multivariate statistical analyses to websites, that of Larson (1996). This study chose sites from a common area, Earth Sciences, and used a cocitation analysis based upon counting pages that link to pairs of chosen sites. Larson produced a plausible subject-based map using multi-dimensional scaling, but his approach does not seem to have spawned any imitators. Björneborn and Ingwersen (2001) have also proposed new web methods that are based, in part, upon bibliometric methodologies, and use ideas from graph theory. Chen et al. (1998) have already studied the web link interconnection between a small subgroup of British universities, using pathfinder network diagrams rather than a group of statistical analysis tools. Thelwall (2001d) has applied non-statistical graphical network techniques to general web domains.

This exploratory study will apply four common multivariate statistical analysis techniques to counts of links between a set of academic websites in order to seek to describe in a useful way the web relationships between them. The sites chosen are from 96 British public university institutions. Britain is an acceptable source because its number of universities gives a data set large enough for meaningful statistical analysis, but just small enough for the  $96 \times 96 - 96 = 9120$  link counts to be obtained. The analysis would not be appropriate for countries with small numbers of

universities, but those with larger numbers could work with an appropriately sized coherent subset. The primary research problem is to ascertain whether the statistical techniques are capable of extracting meaningful information from the data sets, in terms of being able to be mapped to plausible real world phenomena. The secondary problem is to discover whether they can be used to reveal the existence of outliers in the data that can be verified through the identification of a cause. Essentially, both the appropriateness of the techniques and the quality of the data are being tested here. This is expected to be the first step in the much longer project of understanding and successfully exploiting this kind of web link data.

## **Method**

### ***Multivariate statistical analysis***

Multivariate statistical analysis has been used with much success to analyse citation data in Author Co-citation Analysis (White and Griffith, 1982). University websites will be the basic unit of study, replacing the 'author oeuvre' of ACA, making the institution rather than the scholar the object of study. Links will be counted rather than co-backlinks: web pages that contain a link to two given universities. This is because few web pages actually represent research comparable to journal articles, and so it should not be expected that two universities would be linked to on the same page to indicate their collaboration on a piece of work. Rather, such information would be more likely to be conveyed by the two universities linking to each other on the pages belonging to the researchers, research groups or project website. Correlation coefficients for matrices of link counts will be used to compensate for differences in scale. When studying web links, the focus can either be upon the source of those links or their targets. In terms of university websites, the decision is whether to compare universities according to the pattern of linking to other universities, or the pattern of other universities linking to them. This manifests itself mathematically as a choice of whether to calculate correlation coefficients between links to or links from pairs of universities. The two phenomena are related over the spectrum of universities, but not identical. Essentially, one is under the control of the university, reflecting its desired link targets, whereas the other is not under its control and may, for example, reflect the impact of its website in a way formalised by the related web impact factor calculation (Ingwersen, 1998). Both types will be calculated and demonstrated here.

### ***Data collection***

The survey will use AltaVista for raw data, rather than a specialist crawler, because of its relatively large coverage of the Web, recent evidence of the reliability of its results and its Boolean syntax that allows the necessary queries to be accurately stated. The use of a commercial search engine does have a fundamental flaw, however, in that its design is both not under the control of the researcher and opaque for reasons of commercial confidentiality, and so the flagging of the issue of data validity is important.

AltaVista's advanced web queries can be used to find the approximate number of pages in one website that link to another. The AltaVista query is based upon the lexicon of the domain names of web page URLs. The key commands are `host:` and `link:` which restrict matches to domain names and to pages with a link to an URL containing the given text, respectively. For example the following query is intended to

retrieve pages at Aston University ([www.aston.ac.uk](http://www.aston.ac.uk)) that contain a link to the School of Oriental and African Studies ([www.soas.ac.uk](http://www.soas.ac.uk)).

host:aston.ac.uk AND link:soas.ac.uk

For universities with more than one commonly used domain name, extended Boolean expressions can easily be formed. For the Royal Holloway University Medical School ([www.rhums.ac.uk](http://www.rhums.ac.uk) and [www.rbhms.ac.uk](http://www.rbhms.ac.uk)) pages containing a link to the University of Lincoln and Humberside pages ([www.ulh.ac.uk](http://www.ulh.ac.uk), [www.lincoln.ac.uk](http://www.lincoln.ac.uk) and [www.humber.ac.uk](http://www.humber.ac.uk)) the following query would be effective.

(host:rhbnc.ac.uk OR host:rhul.ac.uk) AND (link:ulh.ac.uk OR link:lincoln.ac.uk OR link:humber.ac.uk).

Such queries will nonetheless miss all pages on unknown websites. It is common practice in Britain to use derivative domain names, but this is not ubiquitous. The type of search described above would be inclusive of subdomains such as Birmingham University's School of Computer Science ([cs.bham.ac.uk](http://cs.bham.ac.uk) from [bham.ac.uk](http://bham.ac.uk)), but not of Manchester University's computer centre ([mcc.ac.uk](http://mcc.ac.uk) from [man.ac.uk](http://man.ac.uk)). It is impossible to identify all the non-derivative domain names used by British universities, particularly because commercial names are sometimes used, hence a source of unreliability of the data. The above technique does not work for one British university, St Andrews. This is because its domain name contains a hyphen followed by a word that has a Boolean interpretation: [www.st-and.ac.uk](http://www.st-and.ac.uk), resulting in unreliable performance in AltaVista. As a result of this St Andrews was omitted.

In summary, link counts between 96 British universities were obtained from AltaVista, excluding Glamorgan because of its very limited coverage by AltaVista and St Andrews because of the technical problem described above.

## Results

### *Factor analysis*

A factor analysis (using principal component analysis without rotation) on the correlation coefficients for universities as the target of links produced ten factors. Of these only two were easily attributed to a cause: one for Scottish universities and one for universities in Greater Manchester. Two underlying trends in the data, then, are for the Scottish universities to have a similar spread of backlink sources and the same for Manchester-based universities. It is presumed that these factors identify an underlying tendency for these universities to link more to others in the same group. This is not a necessary conclusion from the existence of a factor in this context: all that can be directly inferred is an underlying tendency to have a similar profile of backlink sources. Tentative suggestions for three of the remaining eight factors were southern universities (no universities with loadings over 0.7), midlands universities (no universities with loadings over 0.7), and Imperial College (Imperial College loading 0.76, others < 0.57). In the former two cases the pattern was not perfect, with many exceptions present.

The factor analysis on the correlation coefficients for universities as the sources of links was also difficult to interpret. A Scottish factor was present, but no Manchester

factor in the nine identified. Explanations were constructed for three of the remaining six factors: Brighton (Brighton loading 0.75, others < 0.49), Luton University (Luton 0.77, others < 0.41), and, tentatively, London universities (no universities with loadings over 0.7). Two further factors were labelled new university factor 1 and new university factor 2 for the predominance of new universities with the higher loadings, although in both cases the loadings were all below 0.7.

In summary, clear associations between Scottish universities and between Manchester universities were present in the data, with a suggestion also of different profiles for new universities, particularly in the pattern of the targets of links, and slight southern and London patterns. Imperial College and the Universities of Brighton and Luton also stood out as possessing unusual underlying characteristics, the former in backlink pages and the latter two in link sources. Luton's unusual behaviour as the source of links is easy to spot in the raw link count behaviour: AltaVista records it as only delivering 21 links to all other UK universities put together. Although magnitude of link counts is not an issue because of the correlation coefficient conversion used, the consequent majority of zeros in the raw data would have an impact. Similarly Brighton was unusual for hosting 12,993 links to Imperial College, 93% of its recorded links. This also accounts for Imperial College's anomalous link target profile. The links were caused by Brighton hosting a mirror copy of Imperial's Free Online Dictionary of Computing, containing a credit link on all pages. The unclassified factors may be anomalies in the data or could represent real phenomena such as large web based collaborative projects between universities, although no evidence was found from web searches for the latter explanation. Perhaps the most surprising result was the failure to identify factors for the federal universities of Wales and London. A factor for Northern Ireland was not expected, however, since interlinking of its two universities would not be measured as a result of the university self-link count being treated as a missing value. In both factor analyses, the largest factor contained the majority of universities with a high loading, perhaps indicating a 'normal' link profile.

### ***Cluster analysis***

Two hierarchical cluster analyses using the between-groups linkage method were conducted, the first for correlation coefficients between link counts of links to universities, the second for links from universities. Clusters were identified loosely from the patterns in order to extract as much suggestive information as possible from the data.

The first cluster analysis grouped all the Scottish universities into two clear, but separate, groups. The strongest grouping was of the five new universities, plus Heriot-Watt. The other grouping included the remaining six traditional universities. There were other clusters that could be interpreted geographically. A Manchester cluster included Salford, UMIST and Manchester Metropolitan University, but not the University of Manchester. There was also a cluster for the high research quality central midlands universities: Birmingham, Loughborough and Warwick. This excluded Aston, however, which is also a traditional university in the same area. The remaining clusters were not dominated by a region, nor were they exclusively, or almost exclusively, of high research or low research institutions.

The second cluster analysis, using links from universities, was not as clear-cut for the Scottish universities. Ten of the twelve were placed in two adjacent clusters, although one of them also contained Birmingham. The two ‘missing’ Scottish universities were Abertay and Napier. There was again a ‘Manchester’ cluster, although this time it also included the University of Northumbria at Newcastle and Liverpool John Moores. There was an additional geographic pair of small clusters, for London. This consisted of Greenwich and City next to the London School of Economics, Westminster and East London, not all of which are in the federal University of London. Many other London-based universities were not in or adjacent to these clusters. There was also a very large cluster of 24 universities, all except one of which were of the higher research traditional variety. The remainder of the groups did not have any clear geographic or research commonality.

Outliers in the data can be identified as the last universities to be associated with others, although there is not a natural cut-off point for how many to include. From the first analysis those identified (in order) were Surrey, UWE, Imperial College, Northumbria, Exeter and Brighton. The second analysis gave Brighton, Royal Holloway, Open, Sheffield Hallam and Abertay, and Surrey.

**Multi-Dimensional Scaling (MDS)**

Two-dimensional graphs were plotted for both matrices of correlation coefficients, using multi-dimensional scaling. Figure 1 shows a graph of the results for correlation coefficients between profiles of links to universities. The more isolated universities are labelled, as are the Scottish, a Manchester group and a cluster of exclusively new universities. Perhaps clearest in the graph is the presence of outliers: universities with profiles of links to them that do not appear to be similar to that of any other. Dimension 1 is almost a research indicator, with new universities predominantly on the left and old universities dominating the right.

Figure 1: *Multi-Dimensional Scaling applied to correlation coefficients between profiles of links to British universities*

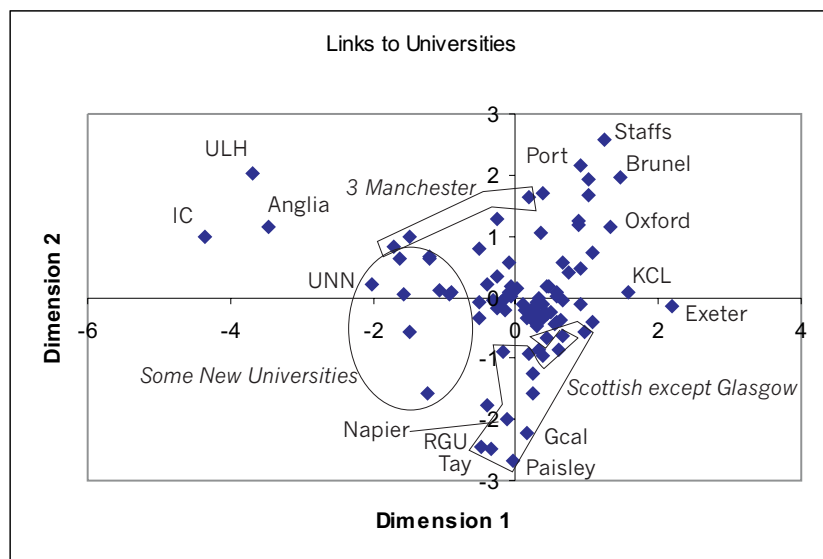
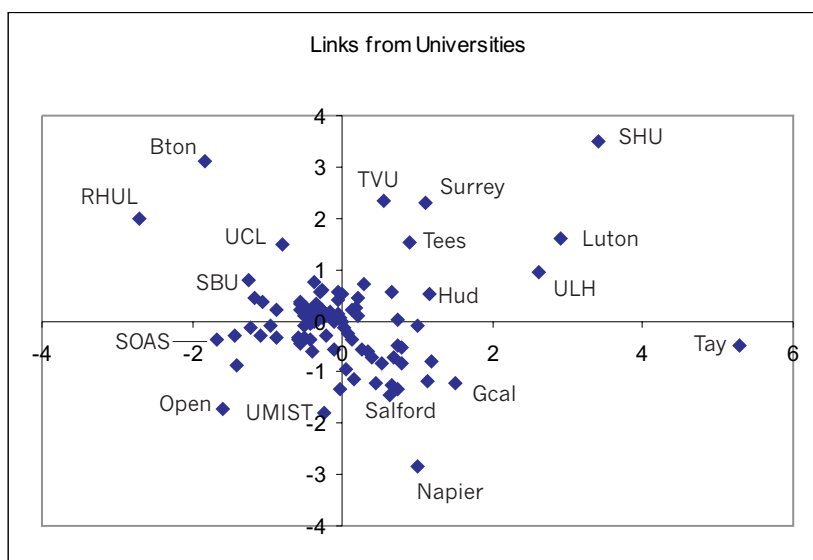


Figure 2 shows results from the second correlation matrix. Isolated universities are again visible but this time it was not possible to plot a complete Scottish group on the graph. Neither of the dimensions corresponds to geographic nor research factors.

Figure 2: *Multi-Dimensional Scaling applied to correlation coefficients between profiles of links from British universities*



The principle outcome of the multidimensional scaling exercise has been in the identification of possible outliers, the identification of groupings being very difficult.

**Maximal spanning trees**

Maximal spanning trees were calculated for both sets of data. These are closely related to pathfinder networks, but are more appropriate for large networks because they contain fewer links (Chen, 1999). Essentially, the diagrams are constructed by repeatedly removing links associated with the smallest weighting but not removing any link that would result in the diagram splitting into two separate halves, stopping when no further links can be removed. The principle advantage of this kind of diagram is its potential to create a meaningful picture from a large quantity of data. Its main drawback is that the simplification necessary to achieve a visually comprehensible diagram can lead to a degree of arbitrariness. Principally, the decision to include one link instead of another can be made based upon a marginal difference between their weightings, which can have a knock-on effect. The technique, therefore, will be strongest when there are large clear contrasts in the link weightings so that link choices do represent significant facets of the data set.

Figures 3 and 4 show a much more marked identification of geographically similar regions than the other techniques. To partially resolve the focus vs. content problem, names are replaced with abbreviations based upon the domain name. See <http://www.scit.wlv.ac.uk/ukinfo/> for a geographic map of all UK universities and information about website names.

Figure 3: A maximal spanning tree for correlation coefficients between profiles of links to British universities

Only the connections between universities are significant, not their actual location in two-dimensional space. New universities are underlined and relatively coherent geographic groupings identified. The abbreviations used are derived from the Internet domain names (e.g. Man: www.man.ac.uk) and so the identity of any unknown universities can be ascertained by visiting their website.

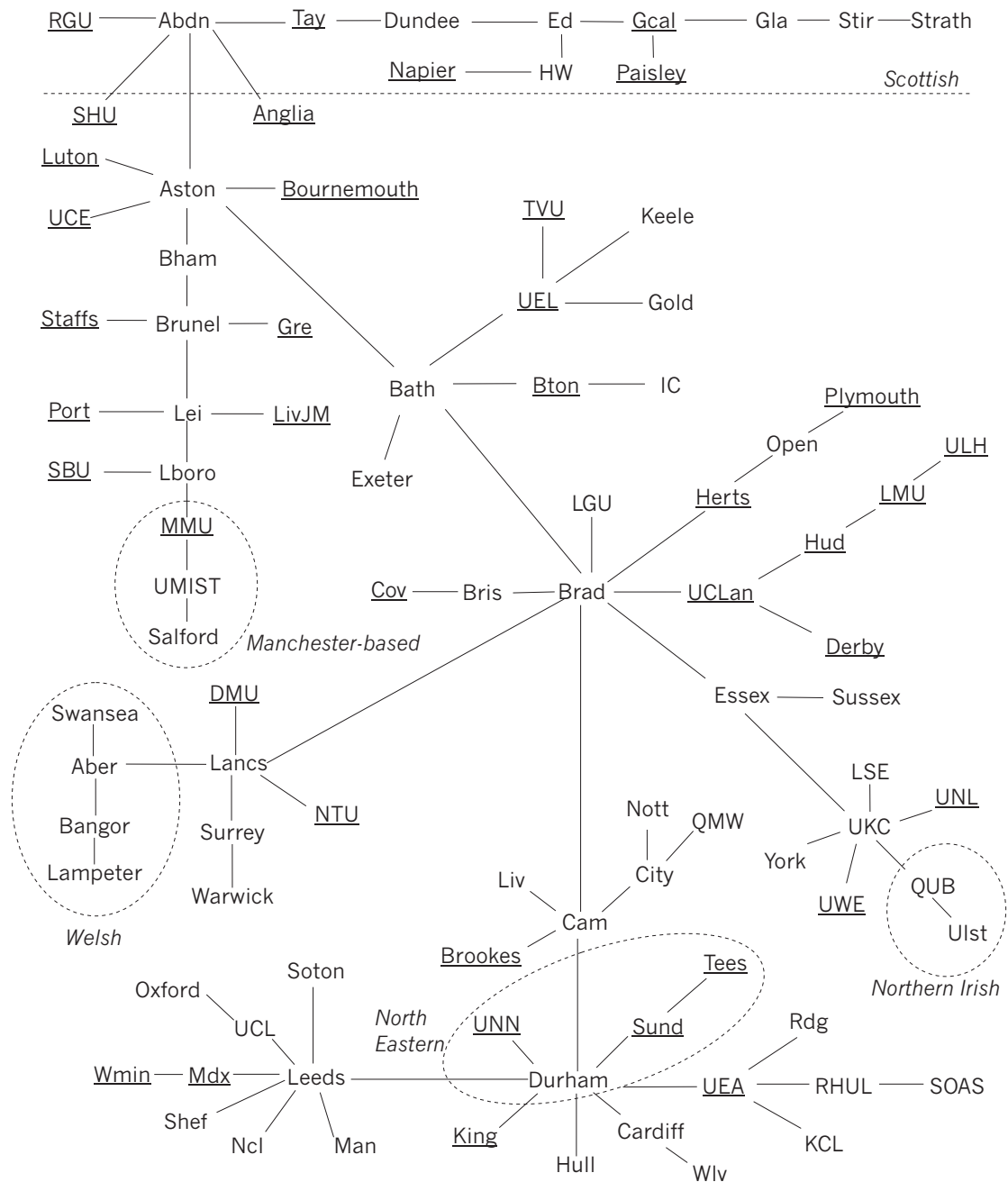
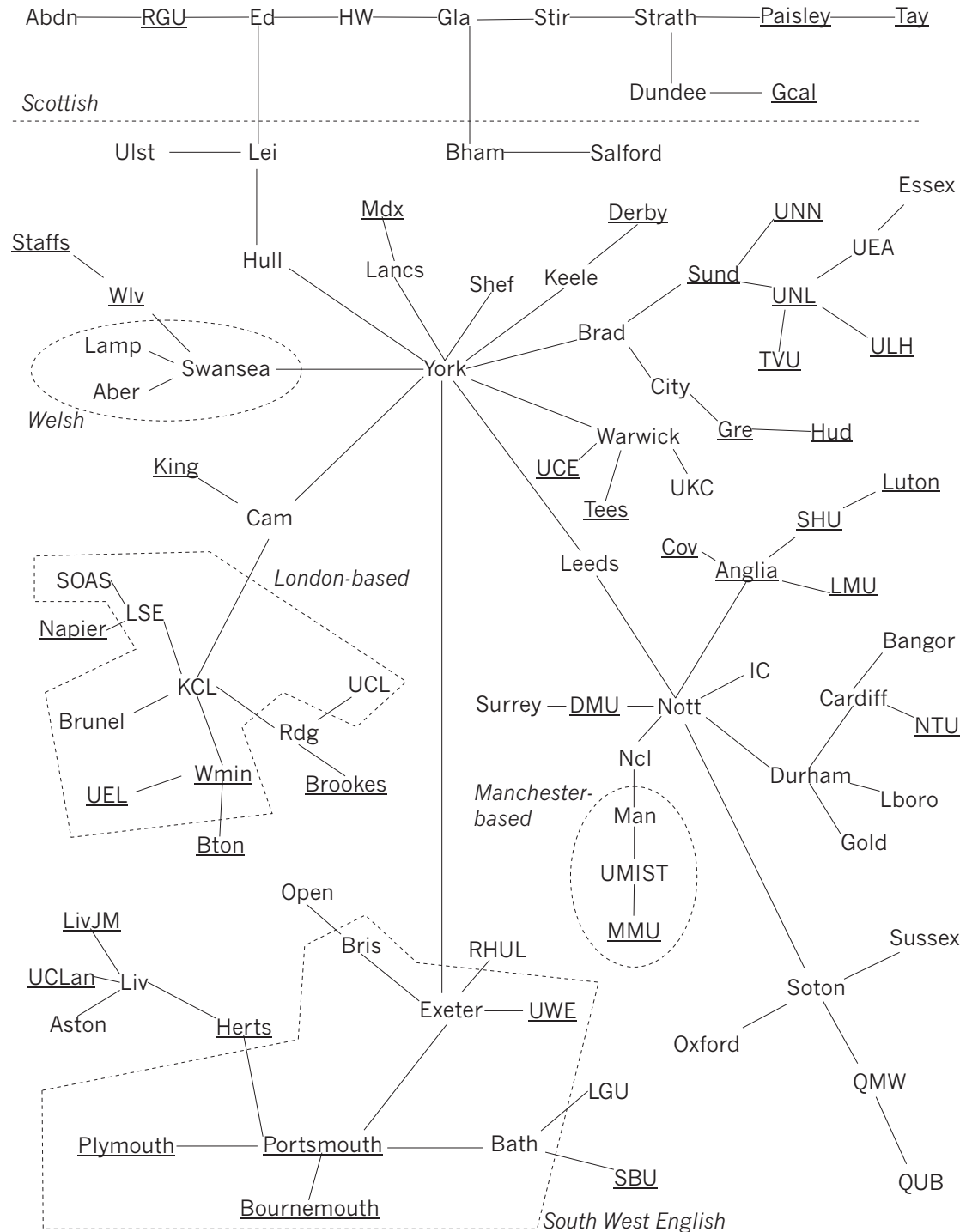


Figure 4: A maximal spanning tree for correlation coefficients between profiles of links from British universities

Only the connections between universities are significant, not their actual location in two-dimensional space. New universities are underlined and relatively coherent geographic groupings identified.



## **Discussion**

### ***The identification of groupings***

The four techniques applied all produced some useful information, although much of it was not clear-cut. The fundamental difference between a set of universities and the scientific field of an ACA exercise was clearly evident in the difficulty in identifying coherent groups. Factor analysis was useful in identifying subgroups with commonality, but some of the factors were hard to relate to real world phenomena. Cluster analysis, factor analysis and maximal spanning trees were able to suggest logical groupings of universities. In terms of the group of universities analysed, some useful information was obtained. A tendency for Scottish universities to form a coherent group was identified and a looser connection for the Welsh, but there was a lack of an identifiable communality between the members of the federal University of London. The only other persistent regional grouping was of the Manchester universities, usually excluding the University of Manchester. There was also a little evidence of grouping along research quality lines. The reliability of the groupings is certainly open to question, although those that recur in different perspectives and techniques inspire more confidence. In terms of the main research question of this paper, this both gives reassurance that the data used is capable of revealing plausible underlying trends, and also the caution that further methodological development is necessary if more useful patterns are to be found.

### ***The identification of outliers in the data***

The secondary research question has been successfully answered, with all of the techniques except maximal spanning trees able to identify potential outliers in the data set, many of which were verified by identifying the causes of the anomalies in the relevant websites. Because of the ease with which large collections of web pages can be created and transported, there is a need to identify and minimise the effect of such anomalies in order to produce more reliable web link counting results (Thelwall, 2002), and so this is a valuable discovery. The outliers produced by factor analysis were more believable than those produced by the other techniques, due to their forced dimension reduction. These may nevertheless prove useful to suggest candidates for data cleansing if further procedures are designed to verify whether a genuinely unusual phenomenon, such as a mirror site, is present, or whether the apparent isolation is an artefact of the statistical technique. The results of a statistical investigation of this kind of data that was preceded by data cleansing may well be revealing.

### ***The underlying structure of the data and the applicability of the techniques***

If the patterns discovered in the web links are genuine then it follows that the underlying structure cannot be easily represented in less than three dimensions, two for geographic location and one for research. For this reason, cluster analysis and standard MDS would be unlikely to be successful at representing university interlinking in two dimensions. A three-dimensional version of MDS may, however, be more successful. Pathfinder analysis or maximal spanning trees would suffer from the same confusion between the two competing orthogonal components in the data. This is not a display dimensionality issue; it is the result of building a model upon a series of binary decisions.

Factor analysis can cope with multidimensionality, but not with the graduated changes implicit in both the research and geographic functions. It could be expected to pick up clustering in the data, in either component, such as a group of universities with similar research standing, or a geographically close body. Its success at this, even for 'cleansed' data, would be limited by the mathematical confusion caused by the more gradually changing relationships present.

## Conclusion

The statistical techniques used have pointed to the underlying data structure being determined by both geography and research. The actual link counts also revealed large anomalies with identifiable causes. The conclusion was reached that although patterns were present in the data, none of the techniques covered were fully adequate to extract them, although all could be expected to provide partial results in certain situations. Further methodological developments are therefore needed. Due to the combination of the anomalies in the data with the graduated research and geographic components, it is likely that a combination of approaches will be necessary to mine more trends from the data. One possible approach is to have an initial data cleansing phase, where anomalies are identified and dealt with, followed by pattern extraction, perhaps using different techniques. The pattern extraction could also be undertaken in two phases, one for research and one for geography. Ultimately, if both of these were successful, the residual trends would point to patterns that would perhaps be the most interesting of all, reflecting less-readily identifiable phenomena.

The results in this paper have served to highlight the issues that must be resolved in order to mine university-university web link count data and it is hoped that the results will form a stepping-stone for the creation of a set of more effective techniques.

## References

- Bar-Ilan, J. (1999). Search engine results over time – a case study on search engine stability. *Cybermetrics*, 2/3. Available: <http://www.cindoc.csic.es/cybermetrics/articles/v2i1p1.html>
- Bar-Ilan, J. (2001). Data collection methods on the web for informetric purposes - a review and analysis. *Scientometrics*, 50(1), 7-32.
- Björneborn, L. and Ingwersen, P. (2001). Perspectives of webometrics. *Scientometrics*, 50(1), 65-82.
- Chen, C., Newman, J., Newman, R. and Rada, R. (1998). How did university departments interweave the Web: a study of connectivity and underlying factors. *Interacting with computers*, 10(4), 353-373.
- Chen, C. (1999). Visualising semantic spaces and author co-citation networks in digital libraries. *Information Processing and Management*, 35(3), 401-420.
- Cronin, B. (2001). Hyperauthorship: a postmodern perversion or evidence of a structural shift in scholarly communication practices? *Journal of the American Society for Information Science & Technology*, 52(7), 558-569.
- Davenport, E. and Cronin, B. (2000). The citation network as a prototype for representing trust in virtual environments. In: Cronin, B. and Atkins, H.B. (eds.). *The web of knowledge: a festschrift in honor of Eugene Garfield*. Metford, NJ: Information Today Inc. ASIS Monograph Series, 517-534.

- Egghe, L. (2000). New informetric aspects of the Internet: some reflections – many problems. *Journal of Information Science*, 26(5), 329-335.
- Harter, S.P. and Ford, C.E. (2000). Web-based analyses of e-journal impact: approaches, problems, and issues. *Journal of the American Society for Information Science*, 51(13), 1159-1176.
- Ingwersen, P. (1998). The calculation of Web Impact Factors. *Journal of Documentation*, 54(2), 236-243.
- Larson, R.R. (1996). Bibliometrics of the World Wide Web: an exploratory analysis of the intellectual structure of cyberspace. ASIS 96. Available: <http://sherlock.berkeley.edu/asis96/asis96.html>.
- Rousseau, R. (1999). Daily time series of common single word searches in AltaVista and NorthernLight. *Cybermetrics*, 2/3. Available: <http://www.cindoc.csic.es/cybermetrics/articles/v2i1p2.html>
- Smith, A.G. (1999). A tale of two web spaces: comparing sites using Web Impact Factors. *Journal of Documentation*, 55(5), 577-592.
- Snyder, H. and Rosenbaum, H. (1999). Can search engines be used for web-link analysis? A critical review. *Journal of Documentation*, 55(4), 375-384.
- Thelwall, M. (2000). Web Impact Factors and search engine coverage, *Journal of Documentation*, 56(2), 185-189.
- 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). The responsiveness of search engine indexes. *Cybermetrics*, 5(1). Available: <http://www.cindoc.csic.es/cybermetrics/articles/v5i1p1.html>
- Thelwall, M. (2001c). Results from a Web Impact Factor crawler, *Journal of Documentation*, 57(2), 177-191.
- Thelwall, M. (2001d). Exploring the link structure of the Web with network diagrams, *Journal of Information Science* 27(6), 393-402.
- Thelwall, M. (2002). Sources of links for WIF Calculations, *Journal of Documentation*, 58(1), 60-72.
- White, H.D. and Griffith, B.C. (1982). Author co-citation: a literature measure of intellectual structure. *Journal of the American Society for Information Science*, 32(3), 163-172.