Link and Co-inlink Network Diagrams with URL Citations or Title Mentions[1]

Mike Thelwall, Pardeep Sud, David Wilkinson
Statistical Cybermetrics Research Group, School of Technology, University of Wolverhampton, Wulfruna Street, Wolverhampton WV1 1LY, UK.
E-mail: m.thelwall@wlv.ac.uk, p.sud@wlv.ac.uk, d.wilkinson@wlv.ac.uk
Tel: +44 1902 321470, +44 1902 528549, +44 1902 321452 Fax: +44 1902 321478

Webometric network analyses have been used to map the connectivity of groups of web sites in order to identify clusters, important sites or overall structure. Such analyses have mainly been based upon hyperlink counts, the number of hyperlinks between a pair of web sites, although some have used title mentions or URL citations instead. The ability to automatically gather hyperlink counts from Yahoo! ceased in April 2011 and the ability to manually gather such counts was due to cease by early 2012, creating a need for alternatives. This article assesses URL citations and title mentions as possible replacements for hyperlinks in both binary and weighted direct link and co-inlink network diagrams. It also assesses three different types of data for the network connections: hit count estimates, counts of matching URLs and filtered counts of matching URLs. Results from analyses of US library and information science departments and UK universities give evidence that metrics based upon URLs or titles can be appropriate replacements for metrics based upon hyperlinks for both binary and weighted networks, although filtered counts of matching URLs are necessary to give the best results for co-title mention and co-URL citation network diagrams.

Introduction

Since the birth of webometrics (Almind & Ingwersen, 1997), one of its main broad methods, link analysis, has undergone theoretical development to become a practical tool in scientometrics and to some extent in the social sciences. Some examples include the Webometrics world universities ranking (Aguillo, Granadino, Ortega, & Prieto, 2006), contributions to EU indicators (Thelwall, 2010) and reports commissioned for the EU Directorate General of Research (e.g., Robinson et al., 2006). Link analysis also has wider uses, such as to investigate the online recommendation of web sites (Bowler, Hong, & He, 2011), the origins of interest in specific organisations (Zhang, Qi, Yang, Shi, & Xu, 2010), the spread of an issue on the web (Ackland & O'Neil, 2011; Introna & Gibbons, 2009; Rogers, 2002, 2005), the online presence of political groups (Ackland, 2005), the structure of web sites (Petricek, Escher, Cox, & Margetts, 2006), the structure of groups of web sites (Adamic & Glance, 2005; Björneborn, 2006; Ortega & Aguillo, 2009; Park, 2003), or the structure of the web itself (Broder et al., 2000). Link analyses have often used web search engines for raw hyperlink data, starting with AltaVista (Ingwersen, 1998), but the last remaining source of hyperlink searches, Yahoo!, was due to cease this service by early 2012 due to its transition to Microsoft's Bing (Yahoo!, 2011b), which has withdrawn most link searches due to overuse (Seidman, 2007). Moreover, Yahoo! ceased support for automatic searches in April 2011 (Yahoo!, 2011a) leaving no remaining automatic source of link data from search engines. Thus, the need for alternatives to hyperlink counts for webometric network diagrams has become urgent.

One alternative to search engines for link data is a personal web crawler. This is a program that is fed with one or more URLs and uses them to start a crawl of an area of the web or a specified set of web sites. Personal web crawlers cannot cover a significant proportion of the web for Webometric data because of the computing resources needed but can gather link data from a defined small area of the web, such as a collection of web sites (Ackland & Gibson, 2004; Rogers, 2010). This means that certain types of link analysis, such as those covering large web sites or many web sites, may be impractical with personal web crawlers. Moreover, co-inlink data (defined below) needs to be

[1] This is a preprint of an article published in the Journal of the American Society for Information Science and Technology © copyright 2012 John Wiley & Sons, Inc. Thelwall, M., Sud, P., & Wilkinson, D. (2012). Link and co-inlink network diagrams with URL citations or title mentions. *Journal of the American Society for Information Science and Technology, 63(4), 805-816*.

gathered from the whole web, if possible, to be more complete. As a result, search engine data seems to be more suitable than personal crawler data for co-inlinks.

There are two main alternatives to hyperlinks that are similar in the sense of identifying connections between web sites. A *URL citation* of web site B by web site A is a page in web site A that contains the URL (or domain name) of web site B but not necessarily a link to it (e.g., "see http://www.bbc.co.uk/news for the latest news"). A *title mention* of web site B by web site A is a page in web site A that contains the name of web site B (e.g., BBC in "the BBC has the latest news"). URL citations are structurally identical to hyperlinks in the sense that they are embedded in one web page and point to another web page. In contrast, title mentions are more general in the sense that the title may be the name of an offline organisation or ambiguous because it is the name of more than one entity.

An important and fundamental difference between title mentions and URL Citations (and hyperlinks) is that title mentions allude to an organisation whereas the latter refers to the organisation's web site. The difference between the two may be minor for online organisations, such as Amazon.com, but significant for organisations that are well known offline, such as universities or other large organisations. Another similar dichotomy is that both hyperlinks and URL citations may be explicit or explicit invitations to navigate to a web site, whereas a title mentions seems much less likely to be a navigational cue. Hence, it should not be assumed that motivations for creating title mentions, URL citations or hyperlinks would be equivalent; there may be significant differences in some contexts.

Link analysis also sometimes uses co-inlinks, which are indirect measures of connectivity. A co-inlink for a pair of web pages or sites A and B is a different web site that contains a link to both A and B. A co-outlink is a page in a different web site that A and B both link to. The links in these definitions are normally hyperlinks, but could also be URL citation "links" or title mention "links", as defined below. Co-inlinks are the web equivalent of co-citations and co-outlinks are the web equivalent of bibliometric coupling (Björneborn & Ingwersen, 2004). Of the two, only co-inlinks have been extensively used for network diagrams. The reason is that good co-outlink data for a collection of web sites relies upon *all* of the web sites maintaining appropriate hyperlinks since all the links counted come from within the sites analysed. In contrast, co-inlink data relies upon links from the rest of the web instead. This is a particular advantage for commercial web sites that have few hyperlinks (Vaughan, Tang, & Du, 2009) even though there is more risk of Spam since the entire web is involved. A co-URL citation of two web sites A and B is a web page in a third web site that contains a URL citation to both A and B (e.g., "See more news at http://www.bbc.co.uk/news and http://www.cnn.com"). Similarly, a co-title mention of A and B is a web page in a third web site that contains a title mention of A and a title mention of B (e.g., "Compare the BBC and CNN news today."). These two reflect indirect connections between web sites or organisations and are potential replacements for hyperlink-based co-inlinks.

This article compares two sets of related metrics for use in two types of network diagram. URL citations and title mentions are compared for direct link network diagrams and co-URL citations and co-title mentions are compared for indirect link network diagrams. Co-title mentions (Vaughan & You, 2010) and URL citations (Stuart & Thelwall, 2006) have previously been used for network diagrams, but have not previously been compared against other data for network diagrams nor evaluated against non-web data. They are potentially useful both for link and colink metrics as a source of method triangulation and as an alternative to hyperlink metrics for automatic searches and also for manual searches when Yahoo! ceases to support hyperlink queries.

Background

This section discusses research using webometric techniques to create network diagrams and some related studies that have used alternatives to hyperlinks for other purposes. A web network normally consists of a set of nodes that are either web sites or web pages, together with a set of connections between the nodes that are identified using hyperlinks or other web data. There are two different types of network: *directed* and *undirected*. In a directed network, the connections between the nodes have a natural direction. If the connections are hyperlinks then the direction would be from the source web site to the target web site. In an undirected network the connections between the nodes have no

direction but serve to connect the nodes in no particular order. Co-inlinks are a type of undirected connection. The type of a web network normally depends upon the type of data used to construct the connections (e.g., co-inlinks or links). Either kind of network can also be *binary* or *weighted*. In a binary network the connections between nodes have no strength so that all connections equal. In a weighted network the connections between nodes have a numerical weight so that some connections can be stronger than others. An example of a weight is the number of links from the source node to the target node. Some networks are naturally binary but it is also possible to convert a weighted network into a binary network by converting all the non-zero weights to 1 (the approach used in the current paper) or by choosing a cut-off value so that connections with weights below the cut-off are removed and connections with strengths above the cut-off value are retained unweighted.

Direct link network diagrams

A hyperlink is a URL embedded in a web page using the HTML anchor tag. It normally associates with text or a picture that the page visitor can click on to navigate to another page. Although designed for navigation, hyperlinks are typically exploited in Webometrics as citation-like inter-document connections. Like citations, hyperlinks are often valued most from the perspective of the targeted document, which presumably has some value or at least a connection to the source document to cause it to be targeted by a hyperlink. Since hyperlinks are valued for their citation-like properties, anything else that functions as an inter-document connection embedded in the source document is a potential alternative and will be referred to as a *direct link* (or just *link* if the context is clear), even when not associated with hyperlinks.

Direct link network diagrams have been created to illustrate the hyperlink relationships within a collection of web sites. In these diagrams, nodes (circles) represent web sites and arrows between nodes represent hyperlinks between them. Sometimes the thickness of the arrows is proportional to the number of hyperlinks (Thelwall & Zuccala, 2008) but arrows can also be all given the same width (Ackland & O'Neil, 2011; Heimeriks, Hörlesberger, & van den Besselaar, 2003; Thelwall, Klitkou, Verbeek, Stuart, & Vincent, 2010). In either case a cut-off may be chosen so that if there are less than a specified number of hyperlinks between a pair of sites then no arrow is drawn.

Whilst direct link networks are sometimes created to reflect the web itself, in webometrics it is more common for them to be employed as a device to investigate communication. For example the web network diagram may be a quick and convenient proxy to identify patterns of collaboration or communication between a set of organisations or individuals based on their web sites (Park, 2010; Park & Thelwall, 2008; Thelwall et al., 2010). A limitation of this approach is that some organisations may wish to hide their connections rather than publicise them. Moreover, businesses may wish to use their web site exclusively as a marketing tool and hence avoid hyperlinks to other web sites altogether (Stuart & Thelwall, 2005). This issue could potentially be ameliorated by using something other than links for inter-document connections but is likely to be impossible to satisfactorily resolve with web-based methods in some cases.

Co-inlink network diagrams

There is a theoretical difference between direct links and co-inlinks. As discussed above, direct links can be used as indicators of collaboration or communication. In contrast, co-inlinks are typically used as indicators of similarity (Chu, He, & Thelwall, 2002; Romero-Frias & Vaughan, 2010; Thelwall & Wilkinson, 2004; Zuccala, 2006). For instance, two companies competing in the same market could expect to have a high co-inlink count (i.e. many web pages simultaneously link to both of them) even if they compete and do not communicate. Similarity may be enhanced in some cases by including topic-related keywords when searching for co-inlinks (Vaughan & You, 2008). Co-inlinks are often the raw data for multidimensional scaling (MDS) diagrams, which indicate similarity by the positions of points in (typically) two-dimensional space but do not explicitly draw the network connections (Chu et al., 2002; Heimeriks & van den Besselaar, 2006; Romero-Frias & Vaughan, 2010; Vaughan, 2006). Other representations used include pathfinder networks (Chen, Newman, Newman, & Rada, 1998), cluster diagrams (Chu et al., 2002) or simple network diagrams (Ortega, Aguillo, Cothey, & Scharnhorst, 2008; Park, 2010). In many contexts, such as academic research, similarity may be a driver of communication. For example, physicists seem more likely to communicate and collaborate

with each other than with historians because of their common knowledge and proximity in conferences and academic departments. Hence, co-inlink data may correlate with direct link data for the same collection of web sites.

Co-inlink counts seem inherently more robust than direct link counts since they are derived from the whole web rather than a small set of web sites and so are less susceptible to anomalous linking by individual web sites. Nevertheless, there are three potential sources of problems. First, if co-inlinks are used as indicators of similarity then there may be an inherent bias caused by the nature of web users and authors. In particular, co-inlinks are likely to be disproportionately large for pairs of sites relating to issues of interest to web authors, such as the web itself, online education or web authoring. This is probably not possible to resolve with web-based methods. Second, if a search engine is used for the co-inlink counts then its results may be unreliable (Bar-Ilan, 2001). Networks are often drawn using the Hit Count Estimates (HCEs) returned by search engines, the figure reported in the results page as being the approximate number of matches for a search. This has been shown to be unreliable to some extent in Yahoo! and Bing for searches with large numbers of results. This is probably because search engines automatically and progressively filter out some matching results because they are duplicates, near duplicates, or come from the same web site as too many previous matches (Gomes & Smith, 2003; Thelwall, 2008b). Thus HCEs from different searches may not be directly comparable since they may be derived from different stages in the filtering process. Alternatively, the full list of matching URLs for each co-inlink search could be retrieved to get exact (filtered) co-inlink counts from each search but this requires extra searches, which is a problem for large networks, and does not work if any count is above the search engine maximum of 1,000. The latter can be partly resolved by the query splitting technique that automatically retrieves additional results beyond the normal 1,000, but at the expense of a greatly increased number of queries (Thelwall, 2008a). Finally, search engines seem sometimes to give results that are wrong, as Liwen Vaughan (personal communication) noticed for Yahoo!'s co-inlink searches.

The quality of co-inlink network data has been assessed in two ways by published studies: author evaluation and external expert evaluation. In both cases, a human judge viewed the diagrams or maps created with the co-inlink data and assessed the extent to which they were reasonable reflections of reality (e.g., the offline similarity of the web site owners) in some way (Heimeriks & van den Besselaar, 2006; Thelwall, 2002; Vaughan et al., 2009). As with the similar bibliometric technique of author co-citation analysis (McCain, 1990), this is probably the best assessment method but has the practical disadvantages of being vague and subjective.

Link data robustness

One drawback of link analysis is that there is often nothing with which to compare the raw data so that its robustness cannot be checked. For example, a network diagram of the hyperlinks between academic web sites may contain anomalies in the form of irrelevant types of hyperlink (e.g., for the web designers' hobbies) but to find these would mean checking all the links individually, a time-consuming task (Harries, Wilkinson, Price, Fairclough, & Thelwall, 2004). Hence there is a need for alternatives to hyperlinks that can be used for method triangulation in order to test the robustness of the results. Two forms of method triangulation have previously been used for link analysis: data sources and data types. Whilst Yahoo! has been the normal source for hyperlink data, web crawlers can also be used to identify hyperlinks. Comparing the two sources can help to identify deficiencies in either. This is practical for link analyses where the sources of the links are within a limited and crawlable fraction of the web. For example, an early study found that the commercial search engine AltaVista had uneven coverage of the UK academic web but that it tended to get more results than the personal crawler SocSciBot (Thelwall, 2001). Another study compared web site size estimates of Google, Yahoo! and MSN (now Bing) for five national audit office web sites with the coverage of the personal crawler Nutch, finding Nutch to give the smallest figures (Petricek et al., 2006). Some research has also compared the results of different search engines (Lewandowski, Wahlig, & Meyer-Bautor, 2006; Uyar, 2009a, 2009b), although not for network diagrams, but this was no longer possible when Yahoo! became the only major search engine usefully reporting hyperlink data. One previous study has attempted to systematically compare data types (Thelwall & Sud, 2011). This compared inlinks to web sites with two other metrics: URL citations of the web site and title mentions

of the organisation owning the web site. The study found inlink counts to correlate significantly with URL citation counts and title mention counts for Yahoo! but found problems with some types of search with Bing. No study has attempted anything similar with link or colink networks, however.

One previous comparative analysis has used a different approach to assess network data. It collected co-inlink data at two points in time using the same methods and compared the multidimensional scaling diagrams produced with them (Vaughan et al., 2009). Other studies have compared network and multidimensional scaling diagrams produced with inlink data with that produced by outlink data (Heimeriks & van den Besselaar, 2006), or have analysed link data in conjunction with networks produced with offline related data (Heimeriks et al., 2003). The purpose in both cases was to gain different insights into the web sites investigated, however, rather than to assess the validity of the different approaches.

Direct link searches with URL citations or title mentions

For direct links from web site A to web site B, the direct URL citation query syntax (Stuart & Thelwall, 2006) for commercial search engines is `"A" site:B` where A stands for the domain name of web site A (or its domain name and path, if it shares a domain name) and B is the domain name of B (or its domain name and path, if it shares a domain name). This query matches all pages in web site B that mention the URL of any web page in web site A. For example, the query `"cnn.com" site:bbc.co.uk` would match any page in the bbc.co.uk web site (i.e., with a domain name ending in bbc.co.uk) in which the text cnn.com was in the page, with or without an associated hyperlink. Since web pages can link to a web site without displaying a visible URL and can display a visible URL without a hyperlink, the direct URL citation is neither more general nor more specific than the direct hyperlink; it is a different inter-document web measure. Research with the related URL citations (queries typically of the form `"A" -site:A`) suggests that direct URL citations will be less numerous than direct hyperlinks, at least in academic environments (Thelwall & Sud, 2011).

A new direct link measure, the direct title mention search, is defined by `"A" site:B` where A is the name of the web site or organisation represented by A, and B is again the domain name of B (or its domain name and path, if it shares a domain name). This query matches pages in web site B that mention web site A without necessarily linking to it or displaying the URL of a page in A. For example `"CNN" site:bbc.co.uk` would match pages in the bbc.co.uk web site that contain the term CNN. Direct title mentions are not more or less general than direct hyperlinks or direct URL citations; they are a different measure. Research with title mentions (queries typically of the form `"A" -site:A`) suggests that they will be less numerous than direct hyperlinks but more numerous than direct URL citations, at least in academic environments (Thelwall & Sud, 2011).

A problem with direct title mentions, and, to a lesser extent with direct URL citations and direct hyperlinks, is that more than one text may be in common use. For example, the BBC can also be referred to as the "British Broadcasting Corporation". This issue can be resolved by using multiple searches, one for each text variant, eliminating duplicates and totalling the remaining results. This can be automated in the Webometric Analyst software (see below). Direct title mentions also potentially suffer from title ambiguity. For example, UCL could refer to University College London or UEFA Champions League. In such circumstances, extra text may need to be added to a title (e.g., "University" in the above case) to ensure that the correct entity is referenced. This has the disadvantage that many appropriate title mentions may not be found.

Co-inlink searches with URL citations and title mentions

The search engine syntax for the new URL citation co-inlink search is `"A" "B" -site:A -site:B`, where A and B are the domain names of web sites A and B respectively. For example, the query `"cnn.com" "bbc.co.uk" -site:cnn.com -site:bbc.co.uk` matches pages outside of the main CNN and BBC web sites that contain both cnn.com and bbc.co.uk in their text, irrespective of the presence of hyperlinks. If A or B share domain names then their path is again used instead of their domain name.

The established title mention co-inlink search (Vaughan & You, 2010) is `"A" "B" -site:A -site:B`, where A and B are the titles of the organisations. Again, in the site: component,

domain names and paths may replace domain names for shared web sites. Also, for the title mention co-inlink search (and the direct title mention search) the query may be duplicated in the case that there are multiple equivalent titles (e.g., UCL and University College London) and expanded if disambiguation is needed (e.g., adding "university" to "UCL" to make the modified title query text: `"UCL" university`).

As for direct links URL citation co-inlinks and title mention co-inlinks are not more or less general than hyperlink co-inlinks. Based upon the same evidence as above title mention co-inlinks should be more numerous than URL citation co-inlinks but less numerous than hyperlink co-inlinks in academic contexts.

Research questions

There are several different methods to construct networks from data derived from web queries. The simplest way is to use the query result HCEs for the strength of connection between a pair of nodes. Alternatively, the number of URLs in the results for a query could be counted. This has the limitation that the result will be artificially low for queries that have more results than the search engine will return. Another alternative is to count the number of URLs matching each query but to filter the results first to remove frequently occurring pages (as suggested by a referee based upon an earlier version of this paper). The rationale for this is that pages matching too many queries may be simple lists of web sites and therefore not "high quality" evidence for connections between web sites. A network constructed by any of the above three methods can be converted into a binary network by reducing all connection strengths that are greater than 1 to 1. This approach is used when the goal is to show where connections exist in a network rather than how strong they are. This gives the first research question.

- Which out of HCEs, URL counts and filtered URL counts give the best results for binary or weighted networks built from direct links or co-inlinks using URL citations or title mentions?

The objective of this study is to assess whether the different direct link searches and the different co-inlink searches that can be automatically calculated with web search engines give broadly similar and valid results for networks. This similarity could be assessed in terms of the most important nodes being the same between methods, or in terms of the overall structural similarity of the networks. This observation drives the following research questions:

- Do the two direct link methods and the two co-inlink methods using URL citations or title mentions give similar rank orders for web sites in binary or weighted networks generated by them?
- Do the two direct link methods and the two co-inlink methods using URL citations or title mentions produce binary or weighted networks with similar structures?

Assessing the validity of the network diagrams produced by the methods is not straightforward. One way would be to show the diagrams produced with the data to experts and ask them whether they are meaningful. An alternative, adopted here, is to compare the results with an external measure that should be related. This gives the following research question.

- Do the two direct link methods and the two co-inlink methods using URL citations or title mentions give rank orders for web sites in binary or weighted networks generated by them that are similar to the rank order of the web sites produced by an external data source of better known validity?

Methods

The overall research design is to compare the results of the different metrics on two unrelated data sets that are relevant to webometrics.

Data

The research questions are general rather than specific to a particular set of web sites. Hence it is likely that the answers will differ between collections of web sites. The approach used here is therefore to assess the results on selected appropriate web networks to suggest the normal likely differences between metrics. There is not an obvious choice of web networks to test, and there is no

register of networks of interest to webometrics from which a random sample could be selected. Instead we selected networks to represent different scales of academic data. The first data set is a collection of 131 UK universities. This represents a network of large web sites. The second data set comprises 49 US library and information science (LIS) departments, as listed in the US & World News 2009 rankings web site (Anonymous, 2009), a source previously used for this purpose in webometric research (Chu et al., 2002). This represents a collection of smaller academic web sites.

For each of the data sets a list of distinctive names and domain names or URLs was created as the basic information needed to make the searches. In most cases organisations had a single unique domain name but in some cases organisations shared a domain name and had to be distinguished by a domain name and path (e.g., http://qcpages.qc.edu/GSLIS/). Identifying names for organisations was more problematic because some had multiple common names, such as a main name and an abbreviation, and some common names were ambiguous (e.g. "School of Library and Information Science"). For the latter issue, titles were sometimes supplemented with other information, such as hosting university name for departments (e.g., "University of Kentucky" "School of Library and Information Science"). This is imperfect because it may generate some incorrect matches and may miss some relevant matches. Hence extensive testing was used to decide upon effective text combinations. For each organisation, alternative common names were sought by trying abbreviations and scanning relevant web pages. For both web networks, the end result of this stage was a text file with a list of relevant title queries and URLs for all web sites. These files were created in the simple Webometric Analyst (http://lexiurl.wlv.ac.uk) input format and were modified versions of files used in a previous publication (Thelwall & Sud, 2011).

Link and co-link networks require many web searches to populate the matrix of connections between the sites. For n sites, n^2-n queries are needed for direct links or (n^2-n)/2 for co-inlinks. The sets of direct link and co-inlink searches were created by Webometric Analyst using new functions added to it for this purpose and with the text files described above as inputs. The files were then used to submit queries to Bing via its Web Search API (Applications Programming Interface) 2.0 through Webometric Analyst on October 14, 2011. The Bing Web Search API was used because this allows automatic searches, which is a practical advantage and common in large scale link analysis. An alternative source of automatic queries to search engines is the University Research Program for Google Search (http://research.google.com/university/search/) but a request to use this received no reply. The Google Custom Search API (http://code.google.com/apis/customsearch/) is another possibility. Although it is designed to search specific sets of web sites rather than the whole web, it can be modified for whole web searches (Liwen Vaughan, personal communication) but testing revealed that its whole web searches gave poor results so it was not used.

The Bing Web Search API results were combined using the Webometric Analyst reporting functions in the case of multiple searches for the same organisations with different names and/or URLs. The same program also converted the results into matrices in the Pajek network format (Nooy, Mrvar, & Batagelj, 2005).

For the `–site:` command of all queries, all URLs were truncated to the main hosting web site (e.g., `–site: qcpages.qc.edu` rather than `–site:qcpages.qc.edu/GSLIS/`) in the belief that this hosting web site would often be too closely tied to the organisation to make it helpful to use its results.

The external source of data for the US LIS departments was the US & World News rakings of 2009. Although this focuses on indicators relevant to students selecting a college, it seems to be a reasonable indicator of academic-related quality. For the UK universities, the totals of the results of the last Research Assessment Exercise (RAE) in 2008 were used (http://www.rae.ac.uk/Results/, accessed July 1, 2011), as previously calculated (Thelwall & Sud, 2011).

Analysis

For the second and fourth research questions Spearman correlations were calculated to compare the rank order of nodes (i.e., universities or departments) in terms of their degree centrality. Using the terminology of social network analysis (Wasserman & Faust, 1994), the degree centrality of a node in an undirected network, such as for the two types of co-inlinks, is the total of all the connections associated with the node. For a directed network, such as for the two direct link networks, indegree

centrality is the most relevant metric and this is the total of all the links pointing to a given node. An alternative metric is outdegree centrality, which is the total of all the links from a given node. A high correlation indicates similar ranks for the nodes from the different sources even if the absolute scores are different. Spearman correlations were chosen because link data is typically skewed. A Bonferroni correction was used to guard against spurious significant results from multiple simultaneous tests. Note that there are other types of centrality than indegree and degree variants introduced above but these measures have been chosen because they are the most commonly used and are more appropriate than others, such as betweenness centrality, that are more dependent on the overall structure of the network.

The standard social sciences matrix similarity QAP correlation test (Hubert & Schultz, 1976; Krackhardt, 1992) was used for the third research question. This is a bootstrapping method that assesses the extent of overlap between two matrices in an unbiased way and is appropriate even if one matrix is denser than the other. This is an important requirement because URL citations and title mentions may not be equally numerous. QAP correlation works by calculating the Pearson correlation between the entries of two matrices and then comparing the result to similar Pearson correlations calculated after the rows and columns of one of the matrices have been randomly permuted. The reported probability value is the chance that the original correlation is lower than a random alternative. Hence, a significant p-value indicates that there is some evidence that the two matrices genuinely correlate.

UCINET was used to calculate the QAP correlations and Webometric Analyst was used to calculate the centrality statistics and to filter the URLs to remove duplicates. For duplicate removal, an arbitrary cut-off frequency of 5 was chosen so that URLs occurring in more than 5 different result sets for the same network were removed from all URL lists for the network.

Results and discussion

Tables 1 to 4 report the correlations between the different data sets and rankings. For the weighted networks reported in Tables 1 and 2, all the metrics correlated significantly with the external sources of information, suggesting that all of the new metrics have some validity for creating weighted networks. Comparing the three different methods (URL counts, filtered URL counts, HCEs) for each of the four metrics, filtered URL counts give higher correlations overall and hence are recommended as the default choice for networks. The performance of the filtered URL list is somewhat counterintuitive for the UK data set because the HCEs are much higher than the number of URLs returned and so the filtering of frequently-occurring URLs might not change the overall results. The reason for filtering working well seems likely to be that more of the frequently-occurring URLs occurred (and hence were filtered out) in less connected pairs of universities. This might occur if the filtered pages were lower quality and did not appear in the results returned for pairs of highly connected universities.

Table 1. Spearman correlations between US & World news ranks and the *weighted* centralities of US LIS department links for 12 different metrics.

Query type*	HCEs	URL counts	Filtered URL counts	Correlation
Co-title			x	-0.743
Co-URL			x	-0.735
Title			x	-0.693
Title		x		-0.659
Title	x			-0.659
URL cite	x			-0.629
URL cite		x		-0.628
URL cite			x	-0.615
Co-title		x		-0.604
Co-title	x			-0.554
Co-URL	x			-0.553
Co-URL		x		-0.549

*Indegree centrality is used for the Title and URL cite metrics and centrality is used for the co-title and co-URL metrics. All correlations are significantly different from 0 at least at $p = 0.05$, including after a Bonferroni correction (Bonferroni correction n=12, gives 0.004 for alpha = 0.05).

Table 2. Spearman correlations between institutional RAE 2008 totals and the *weighted* centralities of UK universities for 12 different metrics.

Query type*	HCEs	URL counts	Filtered URL counts	Correlation
Co-URL			x	0.959
URL cite		x		0.904
URL cite			x	0.903
URL cite	x			0.902
Co-URL		x		0.869
Title			x	0.825
Title		x		0.822
Title	x			0.808
Co-title			x	0.754
Co-title	x			0.708
Co-title		x		0.524
Co-URL	x			0.471

*Indegree centrality is used for the Title and URL cite metrics and centrality is used for the co-title and co-URL metrics. All correlations are significantly different from 0 at least at $p = 0.05$, including after a Bonferroni correction (Bonferroni correction n=12, gives 0.004 for alpha = 0.05).

The correlations for the binary networks in tables 3 and 4 are weaker than those for the weighted networks in tables 1 and 2 but are still mostly statistically significant. Again the filtered URL counts were the best metric. Some of the metrics were not useful because the correlation was zero or because the network was complete, with all nodes being connected to all other nodes.

Table 3. Spearman correlations between US & World news ranks and the *binary* centralities of US LIS department links for 12 different metrics.

Query type*	HCEs	URL counts	Filtered URL counts	Correlation
Co-title			x	-0.718
Title			x	-0.631
URL cite	x			-0.619
URL cite		x		-0.619
Title		x		-0.607
Title	x			-0.607
Co-URL			x	-0.592
Co-URL		x		-0.529
URL cite			x	-0.519
Co-URL	x			-0.421
Co-title		x		-0.329
Co-title	x			0.000

*Indegree centrality is used for the Title and URL cite metrics and centrality is used for the co-title and co-URL metrics. All non-zero correlations are significantly different from 0 at least at $p = 0.05$. After a Bonferroni correction, all are significant except for the last two correlations (Bonferroni correction n=12, gives 0.004 for alpha = 0.05).

Table 4. Spearman correlations between institutional RAE 2008 totals and the *binary* centralities of UK universities for 12 different metrics.

Query type*	HCEs	URL counts	Filtered URL counts	Correlation
URL cite			x	0.878
Co-URL			x	0.875
URL cite		x		0.867
URL cite	x			0.867
Title			x	0.798
Title		x		0.754
Title	x			0.754
Co-title			x	0.696
Co-title		x		-
Co-title	x			-
Co-URL		x		-
Co-URL	x			-

*Indegree centrality is used for the Title and URL cite metrics and centrality is used for the co-title and co-URL metrics. All calculated correlations are significantly different from 0 at least at $p = 0.05$, including after a Bonferroni correction (Bonferroni correction n=12, gives 0.004 for alpha = 0.05).

Tables 5 and 6 support the value of the best metrics from tables 1 and 2, those based upon filtered URL counts, by showing that the centralities of all these metrics correlate significantly. A corollary of this is that all the network diagrams produced should be similar, at least to the extent of having similar most central nodes. This gives good evidence that all four types of measurements, title mentions, URL citations, co-title mentions, and co-URL citations can be useful for producing network diagrams.

Table 5. Spearman correlations between four binary and four weighted centrality metrics for US LIS departments*.

	URL cite filtered	URL cite filtered binary	Title filtered	Title filtered binary	Co-URL filtered	Co-URL filtered binary	Co-title filtered	Co-title filtered binary
URL cite filtered	1.000	0.916	0.726	0.706	0.848	0.749	0.738	0.681
URL cite filtered binary		1.000	0.636	0.614	0.779	0.834	0.642	0.575
Title filtered			1.000	0.967	0.700	0.599	0.949	0.878
Title filtered binary				1.000	0.684	0.601	0.912	0.838
Co-URL filtered					1.000	0.902	0.755	0.691
Co-URL filtered binary						1.000	0.650	0.588
Co-title filtered							1.000	0.907
Co-title filtered binary								1.000

*Indegree centrality is used for the Title and URL cite metrics and centrality is used for the co-title and co-URL metrics. All correlations are significant at p=0.05 with or without a Bonferroni correction (Bonferroni correction n=24, gives 0.002 for alpha = 0.05).

Table 6. Spearman correlations between four binary and four weighted centrality metrics for UK universities*.

	URL cite filtered	URL cite filtered binary	Title filtered	Title filtered binary	Co-URL filtered	Co-URL filtered binary	Co-title filtered	Co-title filtered binary
URL cite filtered	1.000	0.950	0.866	0.849	0.935	0.893	0.751	0.741
URL cite filtered binary		1.000	0.848	0.837	0.931	0.903	0.726	0.725
Title filtered			1.000	0.949	0.859	0.828	0.865	0.815
Title filtered binary				1.000	0.845	0.849	0.802	0.833
Co-URL filtered					1.000	0.949	0.761	0.739
Co-URL filtered binary						1.000	0.706	0.752
Co-title filtered							1.000	0.794
Co-title filtered binary								1.000

*Indegree centrality is used for the Title and URL cite metrics and centrality is used for the co-title and co-URL metrics. All correlations are significant at p=0.05 with or without a Bonferroni correction (Bonferroni correction n=24, gives 0.002 for alpha = 0.05).

Tables 7 and 8 show that most metrics correlate significantly with most other metrics at the whole matrix level, indicate that there will be significant similarities between the networks produced. There are a few exceptions but none of the metrics is an outlier in the sense of correlating with few other metrics. This evidence confirms that all metrics are reasonable for use in making network diagrams.

Table 7. QAP correlations for the filtered metrics for US LIS departments*.

Network	Type	1	2	3	4	5	6	7	8
1	URL cite	1	0.836	0.405	0.208	0.276	0.111	0.177	0.144
2	URL cite binary		1	0.370	0.230	0.267	0.112	0.174	0.126
3	Title			1	0.533	0.571	0.221	0.313	0.219
4	Title binary				1	0.299	0.269	0.232	0.254
5	Co-title					1	0.298	0.389	0.266
6	Co-title binary						1	0.231	0.269
7	Co-URL							1	0.659
8	Co-URL binary								1

* All correlations are significant at p=0.05 with or without a Bonferroni correction (Bonferroni correction n=24, gives 0.002 for alpha = 0.05). Pearson correlations with significance probabilities calculated with a bootstrapping approach: higher correlations are not necessarily more significant with this method

Table 8. QAP correlations for the filtered metrics for the UK universities*.

Network	Type*	1	2	3	4	5	6	7	8
1	URL cite	1	0.434	0.477	0.247	0.216	0.031	0.265	0.086
2	URL cite binary		1	0.356	0.539	0.162	0.064	0.179	0.162
3	Title			1	0.296	0.295	0.038	0.261	0.093
4	Title binary				1	0.076	0.104	0.022	0.138
5	Co-title					1	0.14	0.619	0.240
6	Co-title binary						1	0.095	0.200
7	Co-URL							1	0.289
8	Co-URL binary								1

* All correlations are significant except for the binary network from title searches (number 4) with the weighted co-URL cite network (number 7). If a Bonferroni correction is applied (Bonferroni correction n=24, gives 0.002 for alpha = 0.05) then network 4 also does not correlate significantly with and the weighted co-title network (number 5). Pearson correlations with significance probabilities calculated with a bootstrapping approach: higher correlations are not necessarily more significant with this method

Finally, Table 9 compares the impact of different levels of filtering on the results, by calculating correlations between rankings or RAE 2008 totals and centrality statistics based upon URL counts with the raw data, with the aggressively filtered data used above (URLs occurring at least 5 times being removed) and with less aggressively filtered data (URLs occurring at least 10, 25 or 50 times being removed). Comparing the correlations for the same type of network at the differing filtering levels gives three conclusions. First, filtering has little impact on the directed networks (URL citations and title mentions). Second, aggressive filtering gives a big improvement for the weighted undirected networks (co-title mentions and co-URL citations). Third, any kind of filtering tends to give a small improvement for the binary undirected networks.

Table 9. Correlations between centrality statistics and an external metric for the US LIS departments and the UK universities, showing the difference made by removing URLs that occur at least 5, 10, 25 or 50 times in the results. For title mentions and URL citations, frequent URL removal has little impact but for co-URL citations and co-title mentions it improves the results.

Query type	URL frequencies removed	US+ weighted network correlations	UK++ weighted network correlations	US+ binary network correlations	UK++ binary network correlations
Co-title mentions*	5+	-0.743	0.754	-0.718	0.696
Co-title mentions*	10+	-0.723	0.764	-0.721	0.638
Co-title mentions*	25+	-0.683	0.772	-0.670	0.609
Co-title mentions*	50+	-0.661	0.771	-0.674	0.458
Co-title mentions*	None	-0.604	0.524	-0.329	-
Co-URL citations*	5+	-0.735	0.959	-0.592	0.875
Co-URL citations*	10+	-0.737	0.957	-0.691	0.874
Co-URL citations*	25+	-0.653	0.952	-0.639	0.842
Co-URL citations*	50+	-0.635	0.949	-0.580	0.795
Co-URL citations*	None	-0.549	0.869	-0.529	-
Title mentions**	5+	-0.693	0.825	-0.631	0.798
Title mentions**	10+	-0.666	0.822	-0.735	0.769
Title mentions**	25+	-0.659	0.825	-0.766	0.762
Title mentions**	50+	-0.659	0.824	-0.764	0.759
Title mentions**	None	-0.659	0.822	-0.607	0.754
URL citations**	5+	-0.615	0.903	-0.519	0.878
URL citations**	10+	-0.628	0.903	-0.697	0.903
URL citations**	25+	-0.628	0.908	-0.697	0.871
URL citations**	50+	-0.628	0.909	-0.697	0.871
URL citations**	None	-0.628	0.904	-0.619	0.867

* Indegree centrality metric used, ** Degree centrality metric used
+ Weighted RAE 2008 totals used, ++ US & World News LIS rankings used

Conclusions

In answer to the first research question, the best type of data to use to construct web network diagrams seems to be filtered URL counts (Tables 1 to 4). For the two alternative metrics, URL counts are slightly better than HCEs. For one of the two networks, HCEs and URL counts were significantly inferior for the two types of binary co-inlink networks and so filtered URL counts are particularly recommended for binary co-inlink networks. Filtering out URLs occurring in the results sets of 5 or more queries seems to be an optimal or near-optimal strategy for all types of network. The remainder of the conclusion discusses only networks created with URL lists that are filtered in this way.

In answer to the second and fourth research questions, the results show that, for the two data sets analysed, the rank orders of web site indegrees for weighted and binary URL citation counts and title mention counts correlate significantly with each other and with an external academic-related measure for the web site owners (tables 1 to 6). Similarly, the rank orders of web site degrees for weighted and binary co-URL citation counts and co-title mention counts for co-inlinks correlate significantly with each other and with an external academic-related measure for the web site owners (tables 1 to 6).

In answer to the third research question, the results show that, for the two data sets analysed, the structure of the networks created from weighted and binary URL citation counts and title mention counts correlate significantly with each other (tables 7 and 8). Similarly, the structure of the networks

created from weighted and binary co-URL citation counts and co-title mention counts correlate significantly with each other (tables 7 and 8).

In conclusion, the results suggest that URL citations and title mentions are both reasonable alternatives to hyperlink searches for direct links, although the choice of metric will probably have an influence on the structure of the networks produced. Similarly, co-URL citations and co-title mentions are both reasonable alternatives to hyperlink searches for co-inlinks, although the choice of metric will again probably have an influence on the structure of the network produced. The results do not point to either title mentions or URL citations being the better metric overall, however.

The main limitation of these findings is that only two academic networks were tested and different results may occur in other contexts – particularly for networks that are much larger or much smaller. Nevertheless, the results make it clear that it would be reasonable to use the new searches as replacements for hyperlink variants, although it would be unreasonable to expect them to always give good results. Hence, care and robustness testing is recommended when using any of these data sources. Another limitation is that the research has not assessed qualitatively whether the network diagrams produced are meaningful because this would require subjective judgements. The use of search engines for the data is also an issue because their algorithms can change over time.

An interesting question is whether the URL citation or title mention results are likely to be *better* than the hyperlink results. It seems that title mention searches ought to best reflect connections because creating links seems spurious in the era of Google. Nevertheless, links seem more natural than URL citations and so there seems to be a hierarchy of naturalness, albeit one that is based on weak arguments. The main drawback with title mentions is the need for additional careful human labour to identify appropriate search phrases for each organisation and the problems of ambiguity for titles that cannot always be fully resolved.

The following recommendations are based upon the results of the two experiments.

- For weighted or binary direct link type networks, either URL citations or title mentions can be used, and the data can be based upon HCEs, URL counts or filtered URL counts. The best choice seems to be filtered URL counts, and filtering URLs occurring in 5 or more results sets is a reasonable strategy.
- For *weighted* co-inlink type networks either co-URL citations or co-title mentions can be used, but only filtered URL counts should be used, and the filtering should be aggressive, such as removing URLs that occur in 5 or more results sets. If there is external data for the organisations, then centrality comparisons like those in Table 9 can be used to find the best number for filtering.
- For *binary* co-inlink type networks both co-URL citations and co-title mentions can be used, but only with filtered URL counts, although the level of filtering does not seem to be important. For consistency with the weighted networks, 5 could be used again.

Acknowledgement

This work was supported by a European Union grant by the 7th Framework Programme. It is part of the ACUMEN project (contract 266632). A referee suggested filtering the URL lists and providing results for the filtered URL lists. This improved the conclusions.

References

Ackland, R. (2005). Mapping the U.S. political blogosphere: Are conservative bloggers more prominent? Retrieved February 3, 2006, from http://acsr.anu.edu.au/staff/ackland/papers/polblogs.pdf

Ackland, R., & Gibson, R. (2004). Mapping political party networks on the WWW. *Australian Electronic Governance Conference, 14-15 April, 2004.*, Retrieved December 12, 2009 from: http://voson.anu.edu.au/papers/political_networks.pdf.

Ackland, R., & O'Neil, M. (2011). Online collective identity: The case of the environmental movement. *Social Networks, 33*(3), 177-190.

Adamic, L. A., & Glance, N. (2005). The political blogosphere and the 2004 US election: Divided they blog. *WWW2005 blog workshop*, Retrieved May 5, 2006 from: http://www.blogpulse.com/papers/2005/AdamicGlanceBlogWWW.pdf.

Aguillo, I. F., Granadino, B., Ortega, J. L., & Prieto, J. A. (2006). Scientific research activity and communication measured with cybermetrics indicators. *Journal of the American Society for Information Science and Technology, 57*(10), 1296-1302.

Almind, T. C., & Ingwersen, P. (1997). Informetric analyses on the World Wide Web: Methodological approaches to 'Webometrics'. *Journal of Documentation, 53*(4), 404-426.

Anonymous. (2009). Library and Information Studies. *U.S. News & World Report*, Retrieved February 10, 2011 from: http://grad-schools.usnews.rankingsandreviews.com/best-graduate-schools/top-library-information-science-programs/library-information-science-rankings.

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

Björneborn, L. (2006). 'Mini small worlds' of shortest link paths crossing domain boundaries in an academic Web space *Scientometrics, 68*(3), 395-414.

Björneborn, L., & Ingwersen, P. (2004). Toward a basic framework for webometrics. *Journal of the American Society for Information Science and Technology, 55*(14), 1216-1227.

Bowler, L., Hong, W.-Y., & He, D. (2011). The visibility of health web portals for teens: A hyperlink analysis. *Online Information Review, 35*(3), 443-470.

Broder, A., Kumar, R., Maghoul, F., Raghavan, P., Rajagopalan, S., Stata, R., et al. (2000). Graph structure in the web. *Journal of Computer Networks, 33*(1-6), 309-320.

Chen, C., Newman, J., Newman, R., & Rada, R. (1998). How did university departments interweave the Web: A study of connectivity and underlying factors. *Interacting With Computers, 10*(4), 353-373.

Chu, H., He, S., & Thelwall, M. (2002). Library and information science schools in Canada and USA: A Webometric perspective. *Journal of Education for Library and Information Science, 43*(2), 110-125.

Gomes, B., & Smith, B. T. (2003). Detecting query-specific duplicate documents. *United States Patent 6,615,209*, Available: http://www.patents.com/Detecting-query-specific-duplicate-documents/US6615209/en-US/.

Harries, G., Wilkinson, D., Price, E., Fairclough, R., & Thelwall, M. (2004). Hyperlinks as a data source for science mapping. *Journal of Information Science, 30*(5), 436-447.

Heimeriks, G., Hörlesberger, M., & van den Besselaar, P. (2003). Mapping communication and collaboration in heterogeneous research networks. *Scientometrics, 58*(2), 391-413.

Heimeriks, G., & van den Besselaar, P. (2006). Analyzing hyperlink networks: The meaning of hyperlink-based indicators of knowledge. *Cybermetrics, 10*(1), Retrieved August 1, 2011 from: http://www.cindoc.csic.es/cybermetrics/articles/v2010i2011p2011.html.

Hubert, L., & Schultz, L. (1976). Quadratic assignment as a general data analysis strategy. *British Journal of Mathematical and Statistical Psychology, 29*, 190-241.

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

Introna, L., & Gibbons, A. (2009). Networks and Resistance: Investigating online advocacy networks as a modality for resisting state surveillance. *Surveillance & Society, 6*(3), 233-258.

Krackhardt, D. (1992). A caveat on the use of the quadratic assignment procedure. *Journal of Quantitative Anthropology, 3*, 279-296.

Lewandowski, D., Wahlig, H., & Meyer-Bautor, G. (2006). The freshness of web search engine databases. *Journal of Information Science, 32*(2), 131-148.

McCain, K. W. (1990). Mapping authors in intellectual space: A technical overview. *Journal of the American Society for Information Science, 41*(6), 433-443.

Nooy, W. d., Mrvar, A., & Batagelj, V. (2005). *Exploratory social network analysis with Pajek*. Cambridge: Cambridge University Press.

Ortega, J. L., & Aguillo, I. (2009). Structural analysis of the Iberoamerican academic web. *Revista Espanola de Docomentacion Cientifica, 32*(3), 51-65.

Ortega, J. L., Aguillo, I., Cothey, V., & Scharnhorst, A. (2008). Maps of the academic web in the European Higher Education Area: an exploration of visual web indicators. *Scientometrics, 74*(2), 295-308.

Park, H. W. (2003). Hyperlink network analysis: A new method for the study of social structure on the web. *Connections, 25*(1), 49-61.

Park, H. W. (2010). Mapping the e-science landscape in South Korea using the webometrics method. *Journal of Computer-Mediated Communication, 15*(2), 211-229.

Park, H. W., & Thelwall, M. (2008). Web linkage pattern and social structure using politicians' websites in South Korea. *Quality & Quantity, 42*(6), 687-697.

Petricek, V., Escher, T., Cox, I. J., & Margetts, H. (2006). The web structure of e-government - Developing a methodology for quantitative evaluation. In *Proceedings of the 15th International Conference on World Wide Web (Edinburgh, Scotland, May 23 - 26, 2006)* (pp. 669-678). New York: ACM Press.

Robinson, S., Mentrup, A., Barjak, F., Thelwall, M., Li, X., & Glänzel, W. (2006). *The role of networking in research activities (NetReAct D4.1)*. Bonn, Germany: Empirica Gesellschaft für Kommunikations- und Technologieforschung mbH.

Rogers, R. (2002). Operating issue networks on the Web. *Science as Culture, 11*(2), 191-214.

Rogers, R. (2005). *Information politics on the Web*. Massachusetts: MIT Press.

Rogers, R. (2010). Mapping public Web space with the Issuecrawler. In B. Reber & C. Brossaud (Eds.), *Digital Cognitive Technologies: Epistemology and the Knowledge Economy* (pp. 101-112).

Romero-Frias, E., & Vaughan, L. (2010). European political trends viewed through patterns of Web linking. *Journal of the American Society for Information Science and Technology, 61*(10), 2109-2121.

Seidman, E. (2007). We are flattered, but... *Bing Community*, Retrieved November 11, 2009 from: http://www.bing.com/community/blogs/search/archive/2007/2003/2028/we-are-flattered-but.aspx.

Stuart, D., & Thelwall, M. (2005). What can university-to-government web links tell us about a university's research productivity and the collaborations between universities and government? In *Proceedings of ISSI 2005* (Vol. 1, pp. 188-192).

Stuart, D., & Thelwall, M. (2006). Investigating triple helix relationships using URL citations: A case study of the UK West Midlands automobile industry. *Research Evaluation, 15*(2), 97-106.

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

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

Thelwall, M. (2008a). Extracting accurate and complete results from search engines: Case study Windows Live. *Journal of the American Society for Information Science and Technology, 59*(1), 38-50.

Thelwall, M. (2008b). Quantitative comparisons of search engine results. *Journal of the American Society for Information Science and Technology, 59*(11), 1702-1710.

Thelwall, M. (2010). Webometrics: emergent or doomed? *Information Research, 15*(4), Retrieved July 18 from: http://informationr.net/ir/15-14/colis713.html.

Thelwall, M., Klitkou, A., Verbeek, A., Stuart, D., & Vincent, C. (2010). Policy-relevant webometrics for individual scientific fields. *Journal of the American Society for Information Science and Technology, 61*(7), 1464-1475.

Thelwall, M., & Sud, P. (2011). A comparison of methods for collecting web citation data for academic organisations. *Journal of the American Society for Information Science and Technology, 62*(8), 1488-1497.

Thelwall, M., & Wilkinson, D. (2004). Finding similar academic Web sites with links, bibliometric couplings and colinks. *Information Processing & Management, 40*(3), 515-526.

Thelwall, M., & Zuccala, A. (2008). A university-centred European Union link analysis. *Scientometrics, 75*(3), 407-420.

Uyar, A. (2009a). Google stemming mechanisms. *Journal of Information Science, 35*(5), 499-514.

Uyar, A. (2009b). Investigation of the accuracy of search engine hit counts. *Journal of Information Science, 35*(4), 469-480.

Vaughan, L. (2006). Visualizing linguistic and cultural differences using Web co-link data. *Journal of the American Society for Information Science and Technology, 57*(9), 1178-1193.

Vaughan, L., Tang, J., & Du, J. (2009). Examining the robustness of web co-link analysis. *Online Information Review, 33*(5), 956-972.

Vaughan, L., & You, J. (2008). Content assisted web co-link analysis for competitive intelligence *Scientometrics, 77*(3), 433-444.

Vaughan, L., & You, J. (2010). Word co-occurrences on Webpages as a measure of the relatedness of organizations: A new Webometrics concept. *Journal of Informetrics, 4*(4), 483-491.

Wasserman, S., & Faust, K. (1994). *Social network analysis: Methods and applications*. Cambridge, NY: Cambridge University Press.

Yahoo! (2011a). Web Search APIs from Yahoo! Search. Retrieved April 13, 2011 from: http://developer.yahoo.com/search/web/webSearch.html.

Yahoo! (2011b). When will the change happen? How long will the transition take? , Retrieved July 20, 2011 from: http://help.yahoo.com/l/us/yahoo/search/alliance/alliance-2012.html.

Zhang, Y., Qi, L., Yang, A., Shi, L., & Xu, L. (2010). Investigating spatial distribution of tourist attractions' inlinks: A case study of three mountains. *Science China - Technological Sciences, 53*, 126-133 Suppl. 121 May 2010.

Zuccala, A. (2006). Author cocitation analysis is to intellectual structure as web colink analysis is to…? *Journal of the American Society for Information Science & Technology, 57*(11), 1487-1502.