Interpreting Social Science Link Analysis Research: A Theoretical Framework

Mike Thelwall[1]
School of Computing and Information Technology, University of Wolverhampton, Wulfruna Street, Wolverhampton WV1 1SB, UK. E-mail: m.thelwall@wlv.ac.uk
Tel: +44 1902 321470 Fax: +44 1902 321478

Link analysis in various forms is now an established technique in many different subjects, reflecting the perceived importance of links and that of the web. A critical but very difficult issue is how to interpret the results of social science link analyses. It is argued that the dynamic nature of the web, its lack of quality control and the online proliferation of copying and imitation mean that methodologies operating within a highly positivist, quantitative framework are ineffective. Conversely, the sheer variety of the web makes qualitative methodologies and pure reason very problematic to apply to large-scale studies. Methodology triangulation is consequently advocated, in combination with a warning that the web is incapable of giving definitive answers to large-scale link analysis research questions concerning social factors underlying link creation. Finally, it is claimed that whilst theoretical frameworks with which to guide research are appropriate, a Theory of Link Analysis is not possible.

Introduction

The significance of the web for social and economic life of humans in the developed nations needs no introduction; it is well-rehearsed in academia. There is consequently an urgent need to understand the web and to explore the potential new types of knowledge that it may yield (Davenport & Cronin, 2000). A key feature of the web is the ability for pages to interlink. Link analysis is performed in very diverse subjects, from computer science and theoretical physics to information science, communication studies and sociology, as briefly reviewed below. This is a testament both to the importance of the web and to a widespread belief that hyperlinks between web pages can yield useful information. The different subjects have all contributed valuable insights into links and have exploited them for different purposes.

In information science, the potential for link analysis was recognised when the commercial search engine AltaVista released an interface that allowed users to conduct various types of searches for pages containing links. Researchers with experience of citation analysis (Borgman & Furner, 2002) were quick to point to the possibility for analysing web data with established citation techniques (Ingwersen, 1998; Larson, 1996; Rodríguez i Gairín, 1997; Rousseau, 1997), leading to the hope that the web would allow scholarly inter-document connections that were weaker than citations to be investigated easily on a large scale for the first time (Cronin, 2001). The term 'webometrics' was subsequently coined for the quantitative analysis of web-related phenomena, including links, from an information science perspective (Almind & Ingwersen, 1997), in an article that also laid some theoretical foundations for the new field. Subsequently, webometrics has analysed search engine results (Bar-Ilan, 2001) and web page changes over time (Bar-Ilan & Peritz, 2004; Koehler, 2004), in addition to links. Outside of information science there is one further named social science type of link analysis, hyperlink network analysis (Park, 2003), which is part of communication studies. In other social science web research, link analysis has tended to be embedded into broader investigations (e.g., Foot, Schneider, Dougherty, Xenos, & Larsen, 2003; Hine, 2000) or conducted in a one-off fashion (e.g., Vreeland, 2000).

Although social science link analysis research is now established, particularly in information science (Thelwall, Vaughan, & Björneborn, 2005), and much has been written about methodologies, there is no unanimity concerning the question of how to *interpret* link analysis research results. The question of interpretation has been addressed in published studies, but typically from the perspective of individual research questions, rather than from a generalised framework. There is no clearly stated theory or methodology for link count interpretation. As discussed below, many methods have been

[1] *This is a preprint of an article to be published in the Journal of the American Society for Information Science and Technology © copyright 2004 John Wiley & Sons, Inc.* http://www.interscience.wiley.com/

employed to develop interpretations. The interpretation of results is also an issue for the related field of citation analysis, where the problem has lead to repeated calls for a Theory of Citation to aid citation research and citation statistics interpretation (Leydesdorff, 1998). Citation analysis has not yet delivered a generally accepted Theory of Citation. Nevertheless, research continues and results are published without the support of such a theory and so link and citation analysis need theoretical frameworks with which to interpret results. A theoretical framework can be a *replacement* for a theory of citation/link creation, as argued, in the case of citation analysis, by van Raan (1998). The reason for the distinction is that the statistical averaging inherent in counts of citations to a collection of publications means that individual 'unwanted' citation reasons tend not to cause problems. For link analysis, a theoretical framework needs to be wide enough to encompass all social science link analyses, as argued below.

In this paper, link analysis research traditions are briefly reviewed, and then an argument is made for the necessity of interpreting link counts. A range of direct and indirect methods for link count interpretation are then set out and critically analysed, leading to the development of a new generalised theoretical framework. Finally, citation analysis is revisited from the perspective of link analysis, to discuss the lessons that may be learned from the new perspective.

Link analysis research traditions and approaches

There are several link analysis research traditions. In statistical physics, mathematical models of web structure and web growth are created (Barabási, 2002). Physicists mathematically model links in a very abstract sense, divorced from web page content and social context, and on a very large scale. In computer science, the relationship between links and the content of web pages is of interest, typically from an information retrieval perspective. Links have been incorporated in algorithms (e.g. for search engines) to retrieve authoritative information from the web (Arasu, Cho, Garcia-Molina, Paepcke, & Raghavan, 2001; Brin & Page, 1998; Kleinberg, 1999). A second computer science application is web mining, for example using links to help interpret acronyms (e.g., Larkey, Ogilvie, Price, & Tamilio, 2000). In both of these applications, the social context of link creation and the interpretation of link meanings is outside of the main research perspective, although social factors are incorporated into the initial and final stages of research, to motivate the ideas (e.g., why links might help information retrieval) and to evaluate the outcomes (e.g., comparative evaluation of search engines). The key issue is the efficacy of the algorithms designed, and there is currently no perceived need for social interpretations of the meaning of their outputs. The typical computer scientist does not ask what it means for Microsoft's web site to rank number 1 in a search for "Bill Gates", but might ask whether this is what will be most useful for searchers, and how links can be used to ensure that the most useful result is top ranked (Henzinger, 2001). Although some science research does produce outcomes concerning social structures and web use, their objectives are typically highly quantitative and centred on algorithm design (Gruhl, Guha, Liben-Nowell, & Tomkins, 2004; Kumar, Novak, Raghavan, & Tomkins, 2003; Kumar, Raghavan, Rajagopalan, & Tomkins, 1999).

In the social sciences, and information science in particular, links are used in an information-centred way, to find out about the information on the web: its structure/interrelationships and its value (Bar-Ilan, 2004; Björneborn, 2001; Chen, Newman, Newman, & Rada, 1998; Ingwersen, 1998). Links are also used from an actor-centred perspective, to find out about the structure of networks of actors, whether individuals or organisations, including the importance of individual actors (Garrido & Halavais, 2003; Park, Barnett, & Nam, 2002; Van Aelst & Walgrave, 2002). Note that 'actor' in this sense means is a human agent or group of human agents, a different meaning from that used in the highly specialised sense of actor-network theory. From the information-centred and actor-centred social sciences perspectives, the interpretation of results is a central issue because links are a mechanism through which to study underlying phenomena, including actor importance. It is these two perspectives that are addressed in this paper. Table 1 summarises link analysis application areas.

Table 1. Types of link analysis

Discipline	Applications	Object of study	Interpretation needed?
Statistical physics	Web models and web growth models	Abstract networks of links and pages	No
Computer science	Information retrieval, web mining	Algorithms involving links	No
Information science and social science	Networks of actors	Actors creating and targeted by links	Yes
Information science and social science	Networks of information	Information sourcing and targeted by links	Yes

Any social science link analysis research exercise must first collect link data before interpreting it, perhaps by browsing web sites, querying search engines or using a personal web crawler. The many issues involved in data collection are dealt with elsewhere (Bar-Ilan, 2001; Björneborn & Ingwersen, 2001; Thelwall, 2005) and are not rehearsed here. Note, however, that links between web sites are typically the object of interest, with links within a web site being of less interest because they either do not connect different actors, or because they are likely to be primarily for site navigation purposes.

The need to interpret link counts

The list below illustrates the types of links that may be investigated in a research context; most already have been. The list is useful to make the following discussion more concrete.

1. Counts of links to each of a set of web sites (e.g. university web sites, academic department web sites, journal web sites, business web sites, non-governmental organisation web sites). These may be used to compare how well linked to the web sites are, perhaps using terminology such as ‘online impact’.
2. Counts of links from each of a set of web sites. These may be used to compare how heavily the web sites link to the rest of the web, perhaps using terminology such as ‘luminosity’.
3. Counts of links between each pair of sites in a set of web sites. These may be used to identify patterns of interconnectivity between web sites, perhaps using terminology such as ‘online communication’.
4. Counts of links from each of a set of web sites to a given site or domain (e.g. links to .edu).
5. Counts of links to each of a set of web sites from a given site or domain (e.g. links from .com).
6. All of 1-5, replacing the web sites with top level domains (e.g. country domains).
7. All of 1-5, replacing the web sites with collections of web sites (e.g. all universities in a single country).
8. All of 1-5, replacing the web sites with web pages (e.g. academics’ personal home pages).

In a small-scale social science link analysis experiment, it may be possible to describe each link separately, avoiding the need for a specific interpretation activity. In web link research that involves counting or recording links a larger scale, however, a simplified description is needed with which to report the results, a reasonable interpretation. This is a classic research problem that occurs in any situation where information is quantised. The researcher has to accept the fact that the description of the quantised data will be a simplification and will therefore loose information, but must still ensure that the description of the quantised data is accurate, especially in the sense of not giving a misleading impression of the data. In other words it must have *face validity* (Neuendorf, 2002, p115).

Interpretations of link counts and words chosen to describe them must be reasonable, in the sense that they would broadly fit the readers’ perceptions of the data, should they investigate it (Smith, 1999). The importance of this statement is underlined by previous research that has used or suggested a multiplicity of words to interpret link counts, including visibility (Vreeland, 2000), trust (Davenport & Cronin, 2000; Palmer, Bailey, & Faraj, 2000), worthiness to be looked at (Brin & Page, 1998), quality (Hernández-Borges et al., 1999), and topic authority (Kleinberg, 1999). Similarly, counts of links between web sites have been cast as non-geographic proximity measures (Park & Thelwall, 2003), international information flows (Park & Thelwall, 2003), relationships in a network of organizations (Garrido & Halavais, 2003), information exports (Uberti, 2004) and business connections (Park et al., 2002). Moreover, some individual links appear to have no meaning at all, not performing a communication role (Thelwall, 2003). The implication of all of these different but reasonable interpretations is that researchers should not assume how to interpret links, without checking in some way.

An alternative perspective for interpreting link counts is that of the research question, when it becomes a classic validity issue in social science research methodology terminology (Tashakkori & Teddlie, 1998). If links are used to infer actor or information relationships then evidence must be presented to demonstrate that this is reasonable. The rhetorical question for the researcher is, “Am I truly measuring what I intend to measure, rather than something else?” (Tashakkori & Teddlie, 1998, p80).

Interpreting link counts: Direct approaches

Suppose that an experimenter investigates a set of web sites and obtains a set of link counts, perhaps counts of links to each site, from each site, or between each pair of sites. In order to report the findings, the researcher needs to know what language to use to describe the link counts, and what inference to make if, say, one count is double the size of another count. The most direct way of solving this problem is to ask a random selection of link creators why they authored their links. An alternative relatively direct method is to take a random sample of links and categorise them in a way that is relevant and helpful for the research goals. Both of these approaches are direct in the sense of interpreting the links themselves.

Link creator interviews

Although there have been many investigations into motivations for web site creation (Abbott, 2001; Hine, 2000) and link creation within coherent hypertexts, there do not appear to have been systematic author interviews to discover why inter-site links are created in general web pages. There has been one academic-related study of URL use, however. In this research, Kim (2000) has interviewed 15 authors of academic papers to find out why they included URLs in their citations. This was fruitful, revealing differences between URL citations and traditional citations. Nevertheless, interviews have several drawbacks (Kim, 2000).

- Authors have difficulty in remembering why they included URLs.
- There are practical difficulties in finding and interviewing people, which makes it difficult to operate on a large scale.
- The results can be biased if some authors refuse to participate.

For a general web study, there are significant additional problems.

- Link creation in a general context is presumably less memorable than URL citation in a research article (as in Kim’s study), even if the URL citation actually takes the form of link creation.
- Finding the author of any given web page can be difficult.
- Presumably authors of web pages have less investment in their product than authors of papers, so high participation rates would be difficult to obtain.

Geographical scattering alone probably means that direct interviewing of a random sample of authors is impossible. An alternative is to use questionnaires, or conduct a case study style of research and interview a geographically accessible group of authors. For example, Kim’s interviewees were all from Indiana University.

Classification of random samples

The second way to help interpret link counts is to take a random sample and then instigate a classification exercise to identify the most common types of links and to estimate the proportion of each type of link in the full data set. A small-scale study can perhaps classify all links without sampling and reject those falling outside of a relevant category. This is not a practical solution for large-scale exercises, which must select a random sample of links to classify. Statistical techniques can be used to assess the reliability of proportion estimates generated from a random sample. Although this is not a methods-oriented article, the details of the statistics are relevant to a discussion of the effectiveness of the classification approach.

A randomly chosen set of links for classification is called the *sample* in statistical terminology, and the full set of links from which it is drawn is the *population*. Given a random sample of links, relevant categories can be designed and the proportion of links that match each category can be calculated. What is really needed, however, is the proportion of links in each category in the whole population. A standard statistical technique is to compute a 95% confidence interval, which is a range of values that has a 95% chance of containing the real population proportion. For example, if the proportion of links matching a category in the sample was 0.32 then a statistical calculation might show that there was a 95% chance that the proportion of links matching the category in the whole population was between 0.28 and 0.36.

The accuracy of a proportion estimate depends upon the sample size. A classification of 1,000 links will produce more accurate proportion estimates than a classification of 100. This accuracy will manifest itself in the width of the confidence interval: a more accurate estimate means a narrower confidence interval. If a researcher knows in advance how accurate they want their proportions to be, then she or he can choose a large enough sample size to guarantee the necessary accuracy.

A simple formula can be used to estimate the sample size needed for a given accuracy. Suppose that *e* stands for the allowable error: the amount by which the sample proportion can be bigger or smaller than the population proportion (with a 95% chance). Then the formula for the sample size *n* is as follows.

$$n = \frac{0.96}{e^2}$$

The above formula is derived from Neuendorf (2002) (i.e. Neuendorf's formula d, with $z_c = 1.96$, and $p = 0.5$ as a worst case).

To give an example, if confidence intervals of width 4% are desired, then $2e = 0.04$ so $e = 0.02$, and putting these numbers into the equation (see below) a sample size of 2,400 is needed to guarantee a 95% confidence interval of width no more than 4%. The point of this example is to show that very large numbers of links need to be classified to give reasonably accurate estimates of the proportion of links in a category.

$$n = \frac{0.96}{0.02^2} = 2{,}400$$

Link sampling and link page sampling

Web link research shows that a small number of pages tend to attract a very large number of links, and a small number of pages also host a very large numbers of links (Barabási, 2002). This has been described as a 'rich get richer' phenomenon: pages that are the target of many links are disproportionately likely to be targeted by any new links created. Similarly, pages that already host many links are likely to have additional links added to them. This law also applies to whole sites as well as individual pages (Thelwall & Wilkinson, 2003). The mathematical expression of this relationship is a power law or Lotka's law (Lotka, 1926).

A consequence of the highly skewed distribution of links per page is that a random sample of *links* is likely to be most representative of links to highly targeted pages and links from pages hosting many links. A random sample of link source *pages* might therefore give a very different distribution of link types, in which the highly connected pages are much less represented. In order to interpret link counts effectively, a random sampling exercise should be based upon random samples of links rather than link source pages, since the former represent the phenomenon being interpreted. If commercial search engines are used, however, then since these report link source pages rather than links

(Ingwersen, 1998), random samples should be based upon link pages, rather than links. This is a subtle difference, but is a potential source of confusion that should not be ignored.

The confusion between links and link pages, and the implication of the difference for link interpretation has perhaps not been accorded the importance that it deserves in previous link analysis research (Björneborn, 2003). It will not be discussed further, however, because it is not central to the argument in this paper.

Problems with direct approaches

In this section some serious concerns are raised about the use of author interviews and link classification to attribute meaning to links.

Variety, trends and uniformity

If a sample of links is investigated, either by author interviews or by classification, and the results are used to justify an interpretation of an application of links counts then link type uniformity is an issue. For example, if the application is a network diagram and the interpretation is that links represent informal scholarly communication based upon a random sample, then it may be the case that an alternative characterisation would be more appropriate for the links between *some* pairs of sites. Perhaps certain pairs are connected only by recreational links even though informal scholarly communication links predominate within the whole set. It is known that links are used in different ways on the web, for example with considerable differences between academic disciplines (Harries, Wilkinson, Price, Fairclough, & Thelwall, 2004). Similarly, individual sites may be 'famous' for one thing, perhaps an unusual one such as a mathematical document format conversion program or a clickable geographic map of university web sites (Thelwall, 2002).

Web growth models (Barabási, 2002) suggest that imitation is a powerful factor in Web linking. Almost paradoxically, imitation within an online community can promote variety between different communities. For instance one academic field may tend to link to personal home pages whereas another may tend to link to electronic journals (Harries et al., 2004). Of course the availability of useful resources for a given community (e.g., electronic journals) can also be a powerful factor triggering imitation.

In summary, direct approaches based upon random samples of links (and either author interviews or classification) must assume *uniformity of link types across their sample spaces* in order to apply their link count interpretations; this is problematic because of the lack of uniformity across the web. The ideal solution to this problem would be to classify separate samples for each link count used. In other words, if there were 100 link counts then there would be 100 separate random samples, greatly increasing the work needing to be done. Alternatively, it could be accepted that the generic interpretation would not apply in some cases. A possible hybrid theory-driven approach is to classify separate samples for each class of link count (e.g., one sample per site genre), but linking variety makes this unreliable.

Practical issues: Effort and information value

A practical problem with any classification exercise is the sample size necessary to get reasonably accurate answers. The example above showed that 2,400 links must be classified to get a guaranteed accuracy of +/-2% for each category size. Such an exercise is a major undertaking, raising concerns about whether the value of the information gained is sufficient to justify the effort expended. Taking into account the non-uniformity of web links and conducting multiple classification exercises, as discussed above, further pushes the boundaries of practicality. Note that computer science researchers are able to orchestrate massive web classification exercises through automated means (e.g., Fetterly, Manasse, Najork, & Wiener, 2003) but this is unsuitable because of the typically high error rate; classification exercises would have to be conducted to assess the error rates, which would defeat the point of the automation.

The validity of inference about processes underlying link creation

Fundamental statistical problems can arise if links are not the primary object of study but serve as a vehicle to draw conclusions about processes underlying link creation, which is typical of social sciences link research. This is important for all except the most abstract link analysis research. The theoretical perspective for link counting describes desired link properties if link counts are used to infer underlying human behaviour patterns, as follows (Thelwall, 2005).

> All links counted should be created:
> - individually and independently,
> - by humans,
> - through equivalent judgments about the quality of the information in the target page.

These conditions are unlikely to be met because of factors such as imitation between web authors (the rich get richer phenomenon discussed above), the copying of pages or parts of pages, and automatic creation of web pages by server software (Thelwall, 2005). A consequence of a failure of these properties is that the statistical independence of ‘observations’ (i.e. individual links) means that inferential statistics (including confidence intervals, as discussed above) are not valid. Most statistics are robust to some violation of their underlying principles, but the power law/rich get richer property of links means that the violation in linking is unlikely to be minor. Confidence intervals must therefore be interpreted as indicative rather than accurate when used to infer conclusions about processes underlying link creation. In concrete terms, if physicists link to commercial web sites more than mathematicians, this may not mean that online physicists are more commercially orientated than online mathematicians even if standard statistical tests would back up such a conclusion. The difference may be due to imitation within the respective communities rather than a reflection of differing commercial orientations.

Web dynamics

A problem with any non-historical study is that the situation it describes may have changed by the time it is published. This is a particular concern for web research (Levene & Poulovassilis, 2004, pp. 1-15; Leydesdorff & Curran, 2000). Changes can be both predominantly social and technological, and can affect the number of web links, the way in which they are created, and what they are used for. As a result, the accuracy of any web-related results can be challenged on the basis that link creation may have changed since the links were harvested. This is an argument against large-scale very accurate classification exercises: the accuracy may be spurious. Nevertheless, it is not known how much link use changes over time, although there have been claim of at least one fundamental temporal shift: the disappearance of a significant amount of recreational links in UK academia (Wilkinson, Harries, Thelwall, & Price, 2003). This is a methodological *reliability* issue (Tashakkori & Teddlie, 1998) and is fundamental to the web, as with any rapidly changing environment. Web experiments are inherently unreliable, at least for reporting fine-grained results, given the timescale of academic publication.

As a side issue, there is clearly an important need for longitudinal classification studies to assess the temporal dimension of link creation and give more definitive results about the rate of change of phenomena related to link analysis interpretation. Previous investigations have analysed search engine fluctuations (Bar-Ilan, 1999) and developed methods to cope with these (Rousseau, 1999), but have not addressed concerns about the implications of web dynamics for link analysis research.

Using generic descriptions of link counts

After a classification or interviewing exercise, the issue of what the link counts represent can be addressed. The desired outcome of a classification exercise would be that the vast majority of links fall into categories that can be grouped with a general description that is relevant to the research question. For example, the descriptors ‘interpersonal connection’ or ‘recognition of information value’ might be such general descriptions. There is a grey area concerning the proportion that would constitute a ‘vast majority’, given that some links in any situation will inevitably be of undesired types. Nevertheless, the higher the proportion of links that match desired categories, the more confidence could be claimed for the validity of a link count interpretation. This is a difficulty for the validity of interpretations, since even if a minority of the links represent undesired phenomena, this is still a potential source of invalid interpretations for some of the link counts, especially if the undesired types of links are unevenly distributed. This issue is returned to below (see ‘correlation testing’) and can be resolved in some cases by indirect approaches.

Interpreting link counts: Indirect approaches

In addition to directly investigating links to find out why they have been created or the context in which they are used, there are also less direct methods to suggest appropriate link count interpretations. These include the development of a relevant theory, using the results of previous similar investigations for guidance, and conducting correlation tests with other data sources of known value. The indirect approaches discussed are not mutually exclusive: for example any linking theory should be grounded by a sound literature review.

Literature review

A literature review can be a source of evidence about the use of links in a particular context by alluding to previous relevant research. For example, if one study of inter-departmental links found them to be predominantly related to informal scholarly communication, then this would be corroborate a similar conclusion for an analysis of a different set of inter-departmental links. How similar two investigations are will depend upon a number of factors, including date, country, and site types.

Correlation Testing

Statistical correlation tests can be used to assess the commonality between link counts and data with a better-known meaning, following citation and patent analysis practice (Oppenheim, 2000). This is known as *convergent validity* (Tashakkori & Teddlie, 1998, p83). Research measures have commonly been used to compare with link count statistics for links between universities and departments. A significant correlation between two data sources is evidence that there is some commonality, and is suggestive of a connection between the two data types. It is not proof of causation, however, because there may be an underlying factor that explains the surface commonality between the two. On the web, size is a common factor that can produce spurious correlations: large organisations will tend to have large web sites, many links to their sites, many employees and high revenues. A comparison between any two of these is consequently likely to produce a significant correlation. A more meaningful correlation may be obtained if size is factored out.

Note that correlation testing also addresses a problem that direct methods cannot: that of missing links, through (perhaps unavoidably) inadequate sampling methods. Significant correlation statistics can suggest that the source of the link data is not biased in a way that affects the results.

Correlation testing also highlights a relevant distinction between direct and indirect approaches in the interpretation of link counts, that can also be a difference of perspective between (out)link creation motivations and inlink count interpretations. In citation analysis, aggregated counts of citations to a large group of researchers tend to broadly reflect their research productivity, when appropriately counted (van Raan, 2000), irrespective of the motivations for creation of the individual citations. This may be explained by the observation that if enough of the citations (not necessarily the majority) genuinely reflect research 'quality', and the rest of the citations are broadly random (i.e., at least not systematically non-random) then the average citation count will still reflect research quality. This is almost a paradox: the reasons for being, on average, highly cited can be related to research quality, even if the average reason for citing is not (van Raan, 1998). For link counts, the same may apply: the average reason for creating a link may not be the reason for pages or sites to be highly linked to. This is a strong argument for the importance correlation tests.

Using theory or rational argument

Theses about linking could, in principle, be used as a justification for interpretations of link counts. In the generality of academia, *accepted* theories are typically the result of a period of academic discussion, probably involving discipline-specific epistemological processes, such as empirical experiments (sciences) or debate (humanities) (c.f., Kuhn, 1962). This has not taken place in web research and there is no current accepted Theory of Linking, or generally accepted thesis or framework that could be used to interpret link counts.

Rational argument supporting a thesis is an accepted non-empirical academic approach that is particularly used outside of the sciences. Pure reason is therefore a logical option for interpreting link counts. Reason alone does not present a strong argument for web-related phenomena, however, because of the diversity of uses of the web. In fact there is almost a theory of the atheoretical nature of the web, the 'Loose Web Thesis' (Burnett & Marshall, 2002, pp. 2-3), which emphasises variety and lack of order. A factor further undermining rational argument, and favouring empiricism, is the failure of early extreme cyber-utopian predictions (Hine, 2000). This shows the danger in making untested assumptions about web-related behaviour. In concrete terms, a plausible rational argument for why web links might be created in a given situation may be completely wrong because a significant proportion of web links were created for reasons that had not been conceived by the researchers.

The theoretical framework for link analysis interpretation

None of the above sources of knowledge about links are ideal for the reasons mentioned. The direct sources are problematic because of (a) the practical problems in obtaining a large enough sample size, (b) web dynamics rendering the results of any given sample of temporal interest, (c) the rich get richer linking phenomenon creating fundamental problems with using inferential statistics to interpret factors underlying link counts, and (d) the problem that the reason for high link counts can be completely different to the average reason for link creation (see ‘correlation testing’ above). Web dynamics and web diversity also seem to rule out the generation of a Theory of Linking that could serve as a benchmark or starting point for link analysis studies, and also undermine rationalist approaches to interpreting link counts. Correlation testing is an attractive alternative, but is only an indirect source of information about links, and its use does not avoid the problem of web dynamics.

The logical resolution to these problems is the opposite of a Theory of Linking. It is the adoption of a combination of method triangulation together with the acceptance that link analysis results cannot have a high degree of interpretation reliability, particularly in the temporal dimension.

Method triangulation is the application of more than one method for the same objective so that the combination of methods can shed more light than any individual method (Tashakkori & Teddlie, 1998). Ideally the methods applied should have non-overlapping weaknesses. This is impossible for web link data, however, because of the endemic problems of temporality and heterogeneity. Direct link analysis is desirable as one method because of the dynamic and varied nature of the web. This means either classifying or conducting author interviews for a random sample of pages/links. In practice, the former is the more practical. As a result of the factors discussed above, large sample sizes are not advocated, since these would not be able to give reliable confidence intervals for link types (from the perspective of inferring underling creation motives) and because changes over time are likely. A sample size of 160 seems like a reasonable compromise (Thelwall, 2005). Correlation tests, when possible, should be a second method because the connection with a non-web phenomenon is desirable to cover the potential sample bias weakness of direct approaches, and to cover the possibility that the aggregated counts may need a different interpretation to that of the average motivation, as discussed above.

Link analysis and citation analysis

It is interesting to revisit calls for a Theory of Citation in the light of the theoretical framework above. Compared to links, citer motivations must be more stable over time, yet do change. They are relatively homogeneous due to refereeing but are not uniform because of disciplinary and field differences (Hyland, 2000, chapter 2). The same problems that inflict link analysis also influence citation analysis interpretations, but to a lesser extent. Perhaps the main difference is that there has been an influential (and certainly highly cited) theoretical perspective for citation, that of Merton (1973). Even though it is accepted that Merton’s perspective is an oversimplification (e.g., MacRoberts & MacRoberts, 1996), it is still influential and seems to serve as an implicit justification for much current citation analysis. Practitioners of evaluative bibliometrics, whose reports may influence the careers of the scientists that they study, are careful with their methods (Moed, 2002; van Raan, 2000) and can justifiably claim that there is no need for an extended exercise to help interpret their results: their purpose is to serve as an indicator of research quality, and since statistical evidence demonstrates that citation counts are valuable for this purpose, theories of citation are not needed.

A different argument must be made for relational bibliometrics: for example studies that aim to identify relationships between groups of researchers or groups of journals. These do not always apply the means to divine appropriate citation count interpretations, yet the reason for a close relationship between one group and another may be differ between pairs of groups. For example, journal A may cite journal B heavily for methodological issues, whereas journal C may cite journal D because they cover similar topics. This is not addressed by the correlation/aggregation argument that applies to evaluative citation analysis because research quality is not the central issue. Results of relational investigations may be sometimes taken at face value, however, citer motivation studies (e.g., Chubin & Moitra, 1975) being seen as separate from individual investigations, unless combined for a specific purpose (Oppenheim & Renn, 1978). Relational citation analysis without an embedded interpretative component is nevertheless justifiable when the end users of the research are able to apply their own knowledge to the data. For example, a citation (or author co-citation) map of a field could be intuitively interpreted by field experts who may realise the nature of the citations involved, and this could be reported as part of the findings (McCain, 1990). When non-experts interpret citation data, this is more problematic. A solution to this problem that some adopt is to combine the citation data with a qualitative commentary derived from expert knowledge. From the perspective of web linking, however, Theories of Citation do not seem to be a realistic goal because of the issues of variety, uniformity, dynamics discussed above, even if they apply less to citations. At the risk of over theorising, this may be a post-modern phenomena (Foucault, 1969) because of the lack of definitive explanations, time-resistant theories and the unity or definiteness of objects studied (Law, 2002).

Summary

The theoretical framework for link analysis interpretation is summarised below.

1. Link interpretation is required in any link analysis exercise if conclusions are to be drawn about underlying reasons for link creation – e.g. for all social science link analysis, including information science link analysis. An exception would be made for evaluative link analysis (see the discussion of evaluative bibliometrics in the section above) if it could be consistently demonstrated that inlink counts correlate with the phenomenon desired to be evaluated.
2. No single method for link interpretation is perfect. Method triangulation is required, ideally including a direct method and a correlation testing method.
3. Fundamental problems, including the rich get richer property of link creation and web dynamics, mean that definitive answers cannot be given to most research questions. As a result, research conclusions should always be expressed cautiously
4. Extensive interpretation exercises are not appropriate because of the reasons given in point 3 above.

Finally, for the reasons above, a Theory of Linking, at least in the sense of providing a definitive explanation that can be used to interpret link analysis research findings, is not a realistic research goal.

Acknowledgements

This paper was substantially improved by the insightful comments of the reviewers. The work was supported by a grant from the Common Basis for Science, Technology and Innovation Indicators part of the Improving Human Research Potential specific programme of the Fifth Framework for Research and Technological Development of the European Commission. It is part of the WISER project (Web indicators for scientific, technological and innovation research) (Contract HPV2-CT-2002-00015) (www.webindicators.org).

References

Abbott, C. (2001). Some young male website owners: the technological aesthete, the community builder and the professional activist. *Education, Communication and Information, 1*, 197-212.

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

Arasu, A., Cho, J., Garcia-Molina, H., Paepcke, A., & Raghavan, S. (2001). Searching the Web. *ACM Transactions on Internet Technology, 1*(1), 2-43.

Barabási, A. L. (2002). *Linked: The new science of networks*. Cambridge, Massachusetts: Perseus Publishing.

Bar-Ilan, J. (1999). Search Engine Results over Time - a Case Study on Search Engine Stability. *Cybermetrics, 2/3*, 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.

Bar-Ilan, J. (2004). A microscopic link analysis of academic institutions within a country - the case of Israel. *Scientometrics, 59*(3), 391-403.

Bar-Ilan, J., & Peritz, B. C. (2004). Evolution, continuity, and disappearance of documents on a specific topic on the Web: A longitudinal study of 'informetrics'. *Journal of the American Society for Information Science and Technology, 55*(11), 980 - 990.

Björneborn, L. (2001). *Small-world linkage and co-linkage.* Paper presented at the Proceedings of the 12th ACM Conference on Hypertext and Hypermedia, Århus, Denmark.

Björneborn, L. (2003). Personal communication.

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

Borgman, C. L., & Furner, J. (2002). Scholarly communication and bibliometrics. *Annual Review of Information Science and Technology, 36*, 3-72.

Brin, S., & Page, L. (1998). The anatomy of a large scale hypertextual Web search engine. *Computer Networks and ISDN Systems, 30*(1-7), 107-117.

Burnett, R., & Marshall, P. (2002). *Web theory: An introduction*. London: Routledge.

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

Chubin, D., & Moitra, S. (1975). Content analysis of references: adjunct or alternative to citation counting? *Social Studies of Science, 5*, 423-441.

Cronin, B. (2001). Bibliometrics and beyond: some thoughts on web-based citation analysis. *Journal of Information Science, 27*(1), 1-7.

Davenport, E., & Cronin, B. (2000). The Citation network as a Prototype for Representing Trust In Virtual Environments. In B. Cronin & H. B. Atkins (Eds.), *The Web of Knowledge: a Festschrift in Honor of Eugene Garfield.* (pp. 517-534). Metford, NJ: Information Today Inc. ASIS Monograph Series.

Fetterly, D., Manasse, M., Najork, M., & Wiener, J. (2003). A large-scale study of the evolution of Web pages. *Proceedings of the 12th International World Wide Web Conference*, http://www2003.org/cdrom/papers/refereed/p2097/P2097%2020sources/p2097-fetterly.html.

Foot, K., Schneider, S., Dougherty, M., Xenos, M., & Larsen, E. (2003). Analyzing linking practices: Candidate sites in the 2002 US electoral web sphere. *Journal of Computer Mediated Communication, 8*(4), http://www.ascusc.org/jcmc/vol8/issue4/foot.html.

Foucault, M. (1969). *The Archaeology of Knowledge* (A. Sheridan, Trans.). London: Routledge.

Garrido, M., & Halavais, A. (2003). Mapping networks of support for the Zapatista Movement: Applying Social Network Analysis to study contemporary social movements. In M. McCaughey & M. Ayers (Eds.), *Cyberactivism: Online activism in theory and practice* (pp. 165-184). London: Routledge.

Gruhl, D., Guha, R., Liben-Nowell, D., & Tomkins, A. (2004). *Information diffusion through Blogspace.* Paper presented at the WWW2004, New York, http://www.www2004.org/proceedings/docs/1p491.pdf.

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).

Henzinger, M. R. (2001). Hyperlink analysis for the Web. *IEEE Internet Computing, 5*(1), 45-50.

Hernández-Borges, A. A., Macías-Cervi, P., Gaspar-Guardado, M. A., Torres-Álvarez de Arcaya, M. L., Ruiz-Rabaza, A., & Jiménez-Sosa, A. (1999). Can examination of WWW usage statistics and other indirect quality indicators distinguish the relative quality of medical Web sites? *Journal of Medical Internet Research, 1*(1), http://www.jmir.org/1999/1991/e1991/index.htm.

Hine, C. (2000). *Virtual Ethnography.* London: Sage.

Hyland, K. (2000). *Disciplinary discourses: social interactions in academic writing*. Harlow: Longman.

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

Kim, H. J. (2000). Motivations for hyperlinking in scholarly electronic articles: A qualitative study. *Journal of the American Society for Information Science, 51*(10), 887-899.

Kleinberg, J. M. (1999). Authoritative sources in a hyperlinked environment. *Journal of the ACM,, 46*(5), 604-632.

Koehler, W. (2004). A longitudinal study of Web pages continued: a report after six years. *Information Research, 9*(2), 174.

Kuhn, T. (1962). *The Structure of Scientific Revolutions*. Chicago, IL: University of Chicago Press.

Kumar, R., Novak, J., Raghavan, P., & Tomkins, A. (2003). *On the Bursty Evolution of Blogspace.* Paper presented at the WWW2003, Budapest, Hungary, http://www2003.org/cdrom/papers/refereed/p477/p477-kumar/p477-kumar.htm.

Kumar, R., Raghavan, P., Rajagopalan, S., & Tomkins, A. (1999). *Trawling the web for emerging cyber-communities.* Paper presented at the WWW8, Toronto, http://www8.org/w8-papers/4a-search-mining/trawling/trawling.html.

Larkey, L. S., Ogilvie, P. M., Price, A., & Tamilio, B. (2000). *Acrophile: An automated acronym extractor and server.* Paper presented at the The Fifth ACM Conference on Digital Libraries, San Antonio, TX.

Larson, R. R. (1996). *Bibliometrics of the World Wide Web: an exploratory analysis of the intellectual structure of cyberspace.* Paper presented at the the AISS 59th annual meeting.

Law, J. (2002). *Aircraft stories: Decentering the object in technoscience*. Durham, North Carolina: Duke University.

Levene, M., & Poulovassilis, A. (Eds.). (2004). *Web Dynamics*. Berlin: Springer.

Leydesdorff, L. (1998). Theories of citation. *Scientometrics, 43*(1), 5-25.

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

Lotka, A. J. (1926). The frequency distribution of scientific productivity. *Journal of the Washington Academy of Sciences, 16*(12), 317-323.

MacRoberts, M. H., & MacRoberts, B. R. (1996). Problems of citation analysis. *Scientometrics, 36*(3), 435-444.

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

Merton, R. K. (1973). *The sociology of science. Theoretical and empirical investigations*. Chicago: University of Chicago Press.

Moed, H. F. (2002). The impact-factors debate: the ISI's uses and limits. *Nature, 415*, 731-732.

Neuendorf, K. (2002). *The content analysis guidebook.* London: Sage.

Oppenheim, C. (2000). Do patent citations count? In B. Cronin & H. B. Atkins (Eds.), *The web of knowledge: a festschrift in honor of Eugene Garfield* (pp. 405-432). Metford, NJ: Inormation Today Inc. ASIS Monograph Series.

Oppenheim, C., & Renn, S. (1978). Highly cited old papers and the reasons why they continue to be cited. *Journal of the American Society for Information Science, 29*(5), 225-231.

Palmer, J. W., Bailey, J. P., & Faraj, S. (2000). The role of intermediaries in the development of trust on the WWW: The use and prominence of trusted third parties and privacy statements. *Journal of Computer-Mediated Communication, 5*(3).

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., Barnett, G. A., & Nam, I. (2002). Hyperlink affiliation network structure of top web sites: Examing affiliates with hyperlink in Korea. *Journal of American Society for Information Science and Technology, 53*(7), 592-601.

Park, H. W., & Thelwall, M. (2003). Hyperlink analysis: Between networks and indicators. *Journal of Computer-Mediated Communication, 8*(4), http://www.ascusc.org/jcmc/vol8/issue4/park.html.

Rodríguez i Gairín, J. M. (1997). Valorando el impacto de la información en Internet: AltaVista, el "Citation Index" de la Red. *Revista Española de Documentación Científica, 20*(2), 175-181.

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

Rousseau, R. (1999). Daily time series of common single word searches in AltaVista and NorthernLight. *Cybermetrics, 2/3*, 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.

Tashakkori, A., & Teddlie, C. (1998). *Mixed methodology*. London: Sage.

Thelwall, M. (2002). The top 100 linked pages on UK university Web sites: high inlink counts are not usually directly associated with quality scholarly content. *Journal of Information Science, 28*(6), 485-493.

Thelwall, M. (2003). What is this link doing here? Beginning a fine-grained process of identifying reasons for academic hyperlink creation. *Information Research, 8*(3), http://informationr.net/ir/8-3/paper151.html.

Thelwall, M. (2005). *Link Analysis: An Information Science Approach*. San Diego: Academic Press.

Thelwall, M., Vaughan, L., & Björneborn, L. (2005). Webometrics. *Annual Review of Information Science and Technology, 39*.

Thelwall, M., & Wilkinson, D. (2003). Graph structure in three national academic Webs: Power laws with anomalies. *Journal of American Society for Information Science and Technology, 54*(8), 706-712.

Uberti, E. (2004). *Trading flows and internet hyperlinks: A network analysis.* Paper presented at the AoIR 5.0.

Van Aelst, P., & Walgrave, S. (2002). New media, new movements? The role of the Internet in shaping the 'Anti-Globalization' movement. *Information, Communication & Society, 54*(4), 465-493.

van Raan, A. F. J. (1998). In matters of quantitative studies of science the fault of theorists is offering too little and asking too much. *Scientometrics, 43*(1), 129-148.

van Raan, A. F. J. (2000). The Pandora's Box of Citation Analysis: Measuring Scientific Excellence-The Last Evil? In B. Cronin & H. B. Atkins (Eds.), *The Web of Knowledge: A Festschrift in Honor of Eugene Garfield* (pp. 301-319). Medford, NJ: Information Today, Inc. ASIS Monograph Series.

Vreeland, R. C. (2000). Law libraries in hyperspace: A citation analysis of World Wide Web sites. *Law Library Journal, 92*(1), 9-25.

Wilkinson, D., Harries, G., Thelwall, M., & Price, E. (2003). Motivations for academic Web site interlinking: Evidence for the Web as a novel source of information on informal scholarly communication. *Journal of Information Science, 29*(1), 49-56.