Web Manifestations of Knowledge-based Innovation Systems in the U.K.

David Patrick Stuart B.A.(hons)

A thesis submitted in partial fulfilment of the
requirements of the University of Wolverhampton
for the degree of Doctor of Philosophy

January 2008

Signature……..
Date……..

Publication List

Journal Papers:

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.

Stuart, D., Thelwall, M., & Harries, G. (2007). UK academic web links and collaboration – an exploratory study. *Journal of Information Science*, *33*(2), 231-246.

Thelwall, M., & Stuart, D. (2006). Web crawling ethics revisited: Cost, privacy, and denial of service. *Journal of the American Society for Information Science and Technology*, *57*(13), 1771-1779.

Conference Papers:

Stuart, D., & Thelwall, M. (2005). What can university-to-government web links reveal about university-government collaborations? In P. Ingwersen, & B. Larsen (eds.), *Proceedings of the 10th International Conference of the International Society for Scientometrics and Informetrics*: *Vol*. *1*. (pp.188-192). Stockholm: Karolinska University Press.

Stuart, D., & Thelwall, M. (2007). University-industry-government relationships manifested through MSN reciprocal links. In D. Torres-Salinas, & H. F. Moed (eds.), *Proceedings of the 11th International Conference of the International Society for Scientometrics and Informetrics*: *Vol*. *2*. (pp.731-735). Madrid: CINDOC.

Abstract

Innovation is widely recognised as essential to the modern economy. The term *knowledge-based innovation system* has been used to refer to innovation systems which recognise the importance of an economy's knowledge base and the efficient interactions between important actors from the different sectors of society. Such interactions are thought to enable greater innovation by the system as a whole. Whilst it may not be possible to fully understand all the complex relationships involved within knowledge-based innovation systems, within the field of informetrics bibliometric methodologies have emerged that allows us to analyse some of the relationships that contribute to the innovation process. However, due to the limitations in traditional bibliometric sources it is important to investigate new potential sources of information. The web is one such source. This thesis documents an investigation into the potential of the web to provide information about knowledge-based innovation systems in the United Kingdom.

Within this thesis the link analysis methodologies that have previously been successfully applied to investigations of the academic community (Thelwall, 2004a) are applied to organisations from different sections of society to determine whether link analysis of the web can provide a new source of information about knowledge-based innovation systems in the UK. This study makes the case that data may be collected ethically to provide information about the interconnections between web sites of various different sizes and from within different sectors of society, that there are significant differences in the linking practices of web sites within different sectors, and that reciprocal links provide a better indication of collaboration than uni-directional web links. Most importantly the study shows that the web provides new information about the relationships between organisations, rather than just a repetition of the same information from an alternative source. Whilst the study has shown that there is a lot of potential for the web as a source of information on knowledge-based innovation systems, the same richness that makes it such a potentially useful source makes applications of large scale studies very labour intensive.

Table of Contents

1 General Introduction

1.1 Introduction

Innovation, the successful exploitation of new ideas (DTI, 2007a), is widely recognised as essential to the modern economy. Without it the economy would settle into a stationary state with little or no growth (Fagerberg, 2005). It is therefore unsurprising that there has been an increase in the innovation literature in recent years (Fagerberg, 2005) as attempts are made to have a greater understanding about how the innovation process occurs. A central finding of this literature is the recognition that organisations do not work in isolation (e.g., Lundvall, 1992; Gibbons et al., 1994; Etzkowitz & Leydesdorff, 1995), but rather an organisation depends on "extensive interaction with its environment" (Fagerberg, 2005).

Recognising the importance of an economy's knowledge base and the interactions between different kinds of organisation to the innovation process, Potratz and Widmaier (1996) coined the term *knowledge-based innovation system* to refer to a system where efficient interactions between important actors enables greater innovation. Various models have been proposed in recent years to describe the workings of these knowledge-based innovation systems (Leydesdorff & Meyer, 2003): the Triple-Helix model (Etzkowitz & Leydesdorff, 1995), the National Systems of Innovation model (Lundvall, 1992), and the description of the new 'Mode 2' type of knowledge production (Gibbons et al., 1994). Whilst there are differences between each of these models, they each recognise the growing importance of an economy's knowledge base and the interactions between organisations from different sectors.

Although the importance of knowledge-based innovation systems may be recognised, for changes to be made either at an organisational level, or at a national level, it is important to be able to view where the interactions between actors and the relevant knowledge flows are occurring (OECD & Eurostat, 2005). As Rosenberg states, it is:

> *... central to a more useful framework for analysing the innovation process that it should be based on a more sharply delineated road-map of science/technology relationships. That road-map ought, at a minimum, to identify the most influential traffic flows between science and technology. Obviously, such a map cannot at present be drawn.* (Rosenberg, 1994, p. 139).

Whilst it may not be possible to fully understand all the complex relationships involved within knowledge-based innovation systems, within the field of informetrics, the study of the quantitative aspects of information in any form (Tague-Sutcliffe, 1992), bibliometric methodologies have emerged that allows us to analyse some of the relationships that contribute to the innovation process. Bibliometrics is the application of statistical methods to books and other methods of communications (Pritchard, 1969), and through the application of these methodologies we can understand some of the complex relationships between actors: collaboration may be operationalized through co-authored articles and patents, whilst the flows of ideas may be traced through the citations that link together the networks of scientific papers and patents. Such methodologies are limited, however, by the different publishing cultures of the different sectors, a lack of accepted norms within the sectors, limitations in the tools available to collect the co-authorship and citation data, and the time taken in the publication process. As such it is important to investigate new potential sources of information that can add to the existing informetric sources.

The World Wide Web (the web) is one potential new source of information. The possibility of a linked information system allowing us to see real organisational structures was recognised in Tim Berners-Lee's original CERN proposal in 1990 (Berners-Lee & Fischetti, 1999), and as usage of the web has spread throughout the different sectors of society it offers the potential of allowing us to see the interactions between organisations from different sectors, the basis of knowledge-based innovation systems: more informal relationships than those expressed within scientific papers and patents (Wilkinson, Harries, Thelwall & Price, 2003); collaborations that are not necessarily novel, an essential aspect of scientific papers and patents (Meyer & Bhattacharya, 2004); and collaborations that are in progress rather than those that have already finished (Bossy, 1995). The web may be used in a number of ways to investigate the relationships between organisations, e.g., server access logs, invocations of organisational names or research, or inlinks (Thelwall, 2002g). Unsurprisingly, due to the recognised similarities between hyperlinks and citations, which play an important role in bibliometric investigations, the most popular method adopted is link analysis, analysis of the hyperlinks pointing from one web page to another; a deliberate and explicit reference of one page by another.

This thesis documents an investigation into the potential of the web to provide information about knowledge-based innovation systems in the United Kingdom, a country that has already been the focus of many of the link analysis investigations within the academic community (e.g., Thomas & Willet, 2000; Thelwall, 2002a; 2003a). The rest of this chapter looks more deeply at the ideas touched upon here: the nature of knowledge-based innovation systems; traditional bibliometric methods of investigating knowledge-based innovation systems; the potential of the web as a data source; link analysis as an appropriate methodology; and the United Kingdom as an appropriate area of investigation. The chapter then finishes with a discussion of the aims and objectives of the research and a breakdown of the rest of the thesis.

1.2 Knowledge-based innovation systems

Traditionally the term 'linear model of innovation' has been used to refer to the perception that innovation occurs through the discoveries of basic research effecting applied research, which in turn effects the development and production of new technologies; science is perceived as the driving force of the innovation process. The term 'linear model of innovation' has been ascribed to numerous models which have emphasised the linear nature of the innovation process, each of these linear models may be conceptualised in three main parts: the source of innovation, the process of innovation, and the effect of innovation (Edgerton, 2004) (see Figure 1-1). Such models, however, have been widely criticised: the simplistic nature fails to take into consideration the feedback from different sectors (Kline & Rosenberg, 1986), whilst basic science and use-focused technology do not necessarily have to be mutually exclusive (Stokes, 1997). It has also been questioned whether the linear models were ever more than 'straw men' to begin with (Edgerton, 2004); whilst the linear model has often been attributed to Vannevar Bush's *Science: The Endless Frontier*, it seems that such a claim is false (Stokes, 1997; Edgerton, 2004), rather it has been suggested that it was an over-simplification by the spokesmen of the scientific community to communicate their ideas to the public and policy makers (Stokes, 1997). At best, the linear model may be applied to an extremely narrow range of innovations (Fleck, 2004).

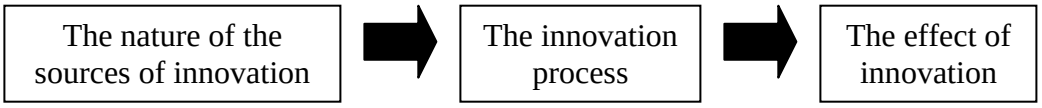

Figure 1-1 The linear model of innovation

Despite the criticisms, and claims that the linear model of innovation is dead (Rosenberg, 1994), the linear model has advantages over its would-be usurpers: the linear model provides an attractive proposition for the creation of simple indicators as it follows that as long as the required resources are put into basic research the economy will get its just deserts (Godin, 2005); and it is well suited to Merton's (1973) norms of science, i.e., universalism, communism, disinterestedness, and organised scepticism. Science as part of an interactive system, reflecting the needs of society is not necessarily palatable to many traditional scientists. Changes in our understanding of the innovation process have been coupled with changes in expectations of the scientific community; it is no longer enough to presume that the benefits of scientific research will happen at some undefined point in the future; there is an expectation that it meets the needs of society today.

The term *knowledge-based innovation system* was coined by Potratz and Widmaier (1996) to refer to the efficient interactions between organisations from different sectors that enable greater innovation by the system as a whole. Whilst Potratz and Widmaier didn't coin the term until 1996 the importance of the interactions between organisations from different sectors of society, i.e., academia, industry and government, as well as with the general public, had already been recognised in a number of works. The term has since been applied, retrospectively, as a broad term to encompass such models. For example, Leydesdorff and Meyer (2003) use *knowledge-based innovation system* as an overarching term to refer to three models that have provoked much discussion in recent years: the Triple-Helix model (Etzkowitz & Leydesdorff, 1995), the National Systems of Innovation model (Lundvall, 1992), and the description of the new 'Mode 2' type of knowledge production (Gibbons et al., 1994). Such a list is by no means exhaustive; the term could also be ascribed to the regional systems of innovation model (Cooke, Uranga, & Etxebarria, 2002) or the finalization model (Schäfer, 1983).

Whilst the different models may be included under one broad title, that is not to say that there aren't fundamental differences between the models: there are differences in which types of organisation are thought to have the leading role (Etzkowitz & Leydesdorff, 2000); and differences in the perceived drivers of the changes seen in the systems (Leydesdorff & Meyer, 2003). However, such differences and the relative advantages and disadvantages of the different models are not the focus of this thesis, instead this thesis focuses on the similarities between the different models, which are encompassed within the term *knowledge-based innovation systems.* Throughout this thesis the term *knowledge-based innovation system* is used to refer to the efficient interactions between organisations from the different sectors that enable greater innovation by the system as a whole.

As well as the more traditional concepts of academic, industrial, and governmental organisations, striving for efficient interactions between the sectors has resulted in hybrid organisations. Industry is seen to be taking on some of the values of universities: sharing knowledge (Etzkowitz & Leydesdorff, 2001), and raising the levels of their training programs and their sponsoring of scientific conferences (Etzkowitz, 2001). Whilst entrepreneurial universities are taking on commercial attributes: exploiting their own research (Lazzeroni & Piccaliga, 2003), and incorporating a managerial culture (Subotzky, 1999). Practices that are

actively encouraged by funding bodies (e.g., Research Councils UK, 2004). Although there are still fears about the effect commercial interests may have on the scientific process (Van Looy, Callaert, & Debackere, 2006), the fiscal realities of the steady state of science (Ziman, 1994) means that there is growing recognition of the need to work with commercial organisations.

Recognition of the importance of the interactions amongst actors from different sectors of society to the innovation process has gone beyond mere academic rhetoric, and is now the focus of government attention. Most notable is the recent inclusion of a new chapter in the latest edition of OECD & Eurostat's (2005) jointly published *Oslo Manual: Guidelines for collecting and interpreting innovation data*, which forms the basis for the European Union's Community Innovation Survey as well as similar surveys in Australia and Canada (OECD & Eurostat, 2005). Although the Oslo Manual recognises the importance of interactions between different organisations, large scale surveys are expensive, require the cooperation of the organisations being surveyed, and are generally overly simplistic. For example, the Community Innovation Survey consists of 28,000 questionnaires in the UK alone, where it has a response rate of 58%. This relatively high response rate reflects the simplistic nature of the questionnaire. Rather than detailing the innovations that have occurred and the flows of knowledge between different organisations, it merely questions whether innovation has occurred, which sectors of society they get their information from, and which sectors they cooperate with (DTI, 2007b). Such standardised simplistic questions encourage their completion and allow comparisons to be made between organisations, sectors, and nations. The information they provide, however, is of limited value in providing a true picture of how innovation occurs and which particular actors are playing key roles in the economy. For certain sections of society more detailed information is obtainable through investigating the organisational connections visible through published documents; the application of bibliometric methodologies (see section 1.3).

1.3 Traditional bibliometric indicators of knowledge-based innovation systems

Whilst information may be collected on the organisational interactions that contribute to the knowledge-based innovation systems through large scale surveys (e.g., OECD & Eurostat, 2005) or through small scale ethnographic studies (e.g., Kogan & Muller, 2006), they are both expensive and reliant on the cooperation of the organisations involved. Informetrics allows the investigation of some of the relationships that contribute to the innovation process through the investigation of bibliographic surrogates. Methodologies that were originally designed for the investigation of science through the inter-document connections of scientific papers (Garfield, 1979), have since been applied to the investigation of technology through the investigation of patent documentation (Narin, 1994), as well as the intersection between science and technology (e.g., Meyer, 2000a). Although scientific papers can provide some insights into the relationships between different actors in the academic community, patents give less indication of the relationships within the technology community, and combinations of the two still leave large areas of knowledge-based innovation systems without any indicators.

Scientific papers provide an opportunity to investigate both broad and narrow conceptions of collaboration. As a broad term, collaboration may be used to refer to the contribution of scientists working towards the furthering of "the common fund of knowledge" (Merton, 1973), and the flow of ideas may be manifested through citations between academic

papers. In the narrow sense, collaboration may be used to refer to those who "contribute directly to all the main research tasks over the duration of the project" (Katz & Martin, 1997), and may be operationalized through co-authorship. Critics, however, point out that there are numerous different reasons for the placement of citations (MacRoberts & MacRoberts, 1989), and that there are multiple scenarios where papers may be co-authored when the parties have not collaborated, or have collaborated without there being a resulting co-authored article (Katz & Martin, 1997; Martin-Sempere, Ray-Rocha, & Garzon-Garcia, 2002). Nonetheless from the perspective of understanding the interactions of knowledge-based innovation systems both co-authorship and citations may be seen as showing a type of interaction between actors. The ability of bibliometrics to provide information about the interactions between actors leads van Raan (2004) to answer Holton's (1978) question of whether science can be measured with a "modest yes"; whilst Katz and Hicks (1996) have suggested that the science citation index can provide systemic indicators of science. This, however, is not enough. The models of knowledge-based innovation systems have increasingly recognised that the measurement of science needs to be viewed in connection with the other sectors of society.

Following the establishment of bibliometric methodologies for analysing science papers there has been a lot of interest in applying similar techniques to patents (Narin, 1994), although the bibliometric methodologies cannot be applied without difficultly. Whilst both science papers and patents can be viewed as networks of papers resulting from intellectual effort, subject to examination, and joined by the inter-document linkages of co-author/inventorship and citation, there are differences between how these are attributed within the different spheres. Whereas the citations within science articles are included by the author (even if this is at the instigation of a referee or editor), patents contain citations by both the examiner and the applicant (Oppenheim, 2000) for different purposes, and whilst there may be similarities between the applicant citations and scientific paper citations, there is a need for further investigations into the mediation process between patent examiner and applicant (Meyer, 2000b). There have also been questions asked about the transfer of co-authorship analyses to patent data; whereas co-authorship is predominantly extramural, technological collaboration is principally intramural (Meyer & Bhattacharya, 2004).

Despite the differences between the two systems there have been a number of investigations into the information each can provide, either in isolation or together, about the nature of the relationship between science and industry: industry-academia collaborations (e.g., Butcher & Jeffrey, 2005), patenting within the academic community (e.g., Webster & Packer, 2001; Cassiman, Glenisson & Van Looy, 2007), the citing of scientific articles by patents (e.g., Vinkler, 1994), and the citing of patents by scientific articles (Glänzel & Meyer, 2003). There are still, however, many interactions that are not likely to appear in any of the bibliometric databases.

Whilst expanding the bibliometric focus to include patents as well as scientific articles provides more information about the interactions between different organisations and the different sectors of society, it is still based on a narrow sample of a formal output. The journal article is not the only published output of the academic community, and not all journal articles are indexed by the bibliographic and citation indexes. Patents are only a by-product of commercial organisations used to protect their intellectual property and as such "patents don't give a complete picture of R & D activity, they only partially describe commercial developments" (Karki & Krishnan, 1997). There are many occasions where it may be felt that the competitive edge provided by commercial developments are thought to be better kept through secrecy rather than patenting (Fisher & Klien, 2003; Lazzeroni & Piccaluga, 2003).

Whilst the collaborations of both science and technology are only partially reflected through the analysis of patents and journal articles, much of the contributions of the government sector are likely to be lost totally. Whilst the work of government research laboratories is likely to appear within patent and bibliometric databases, the role of a government in the effectiveness of a knowledge-based innovation system is far more extensive. In addition, governments provide funding for research at external institutions, implement economic policies that are likely to either encourage or discourage an innovative environment, and provide programmes that encourage organisations to work together. This type of information, as well as much of the informal collaboration and sharing of information between universities, industry, and the public generally, will not be found within the citation and patent databases.

1.4 The web as a source of information on knowledge-based innovation systems

The structural similarities between a directed network of web pages linked by hyperlinks and a directed network of science papers linked by citations was recognised by a number of papers in the mid-nineties (e.g., Bossy, 1995; Larson, 1996), and various terms were suggested for the application of informetric methodologies to the internet and the web: netometrics (Bossy, 1995), internetometrics (Almind & Ingwersen, 1996), cybermetrics (Aguillo, 1997), and webometrics (Almind & Ingwersen, 1997). The similarities between the networks was further emphasised by the term 'sitation' being proposed for one web page being linked to by another, analogous to the traditional citation (Rousseau, 1997). Since then many of the citation analysis techniques originally applied to science papers have been adapted and been applied to the web: the web impact factor is analogous to the journal impact factor (Ingwersen, 1998), co-linked analysis with co-citation analysis (Larson, 1996), and co-link analysis with bibliographic coupling (Thelwall & Wilkinson, 2004).

Of the suggested terms for this new area of investigation, webometrics and cybermetrics have become established within the informetric community, whilst there has been growth in the usage of the term 'webmetrics' within the wider online community (Ingwersen, 2006). Almind and Ingwersen (1997) coined the term webometrics to refer to "research of all network-based communications using informetric or other quantitative measures", although Björneborn (2004) has since redefined it to mean the:

> *...study of the quantitative aspects of the construction and use of information resources, structures and technologies on the Web drawing on bibliometric and informetric approaches. (p. 12)*

Björneborn's inclusion of 'bibliometric' within the definition emphasises the place of webometrics within the larger field of library and information science. More importantly, by applying the term to the web specifically rather than the larger area of the internet, for which he utilises the term cybermetrics, the term resists the tendency for the internet and web to be used interchangeably. The intuitive connection between webometrics and the web, rather than the internet, has led to the popularity of Björneborn's definition and it is in this sense that the term is used throughout this thesis.

At the start of the twenty-first century the web seems a natural place for investigations into the interactions between organisations, at least within the developed countries of the world where organisational use of the web seems near ubiquitous. Initially the preserve of academics, it is now an essential part of business and government for the provision of

information and the promotion of services: in 2005 almost 70% of UK businesses said that they had a web site, which rose to 98% for those with 1000 or more employees (National Statistics, 2006), with 96% of government services reported as being 'e-enabled' at the end of 2005 (Cabinet Office, 2005).

Whilst there are differences in the way the web is utilised within different fields (Kling & McKim, 2000), as well as by the different sectors of society (e.g., Middleton, McConnell, & Davidson, 1999; Shaw, 2001; Musgrave, 2004), the web collocates information that has previously been separate (Cronin & McKim, 1996) allowing connections to be made to works of varying publishable merit, from various sections of society, from anywhere in the world. The web is reshaping the way scholars communicate (Cronin & McKim, 1996) as the differences between formal and informal communication become increasingly blurred (Barjak, 2006) and it is increasingly used as a conversational medium as well as a publishing medium (Cronin, Snyder, Rosenbaum, Martinson & Callahan, 1998). The transformation of the traditional invisible college to a 'cyberspace college' was identified as early as 1994 (Gresham, 1994). It has been suggested that such shifting norms in the way organisations communicate may provide new indicators about the impact and significance of different actors (Cronin, 1999), something that is increasingly applicable as the web is increasingly seen as a respectable source of information (Brown, 2004). Its "dynamic real-time nature" (Ingwersen, 1998) means that it can provide more up-to-date indicators.

Extensive organisational use of and presence on the web does not mean that the web is necessarily a suitable medium for the investigation of the interactions between different organisations, or that such interactions can be linked to the knowledge-based innovation systems that are thought to play such an important role within the economy. Nevertheless it is an important area of investigation, as recognised in a steadily growing number of papers (e.g., Boudarides, Sigrist & Alvious, 1999; Leydesdorff & Curran, 2000; Heimeriks & Van den Besselaar, 2006), recently culminating with a special issue of *Research Evaluation* on web indicators for innovation systems (Katz, 2006). This thesis investigates a link analysis approach to the topic.

1.5 Link analysis

Link analysis is one of the main areas of investigation within webometrics, as well as being the focus of investigations within a wide range of other fields, including mathematics, theoretical physics, communication studies, and sociology (Thelwall, 2004a). Thelwall (2004c) defines the information science approach to link analysis as the adoption and adaptation of "information science techniques for the meta-analysis of documents through investigating inter-document connections" (Thelwall, 2004a, p. 3) with the primary objective being the delivery of useful information (Thelwall, 2004a). The information science techniques were originally designed for the meta-analysis of the inter-document connections of scientific papers: citation-analysis was proposed as a method of mapping the history and structure of science, and identifying core journals, articles and authors within a field (Garfield, 1979); co-citation analysis, bibliographic coupling, and co-word analysis were proposed as methods for the mapping of science (Moed, 2005); whilst co-author analysis enables investigations into the amount of collaboration within a field, which in turn provides an indicator of the maturity of the field (Patel, 1973). These techniques have since been adapted to investigations of the history and structure of technology as well as science, with patents

taking the place of scientific papers as, like science papers, they include citations to earlier literature and details of the inventor and their institutional affiliation (Oppenheim, 2000).

Inter-document connections may be made in a number of ways on the web: through web pages sharing terminology; being linked to by the same document; linking to the same document; or through one document linking to another. Whilst link analysis may be used to refer to any of these inter-document connections (Thelwall, 2004a), or in a broader sense to refer to any of the inter-document connections both on and off the web (Borgman & Furner, 2002), within this thesis, for simplicity, it is used to refer to the analysis of web pages directly connected either by hyperlinks or the closely related URL citation (defined below), unless otherwise indicated. Whereas the sharing of terminology, being linked to by the same document, or linking to the same document are not necessarily a deliberate connection between the two documents on the part of one of the authors, but rather are the construct of the analyst (Leydesdorff, 1998), the placement of a web link or URL citation is a deliberate connection. Therefore it is direct web links and URL citations that are investigated in this thesis to see what they can tell us about the interactions between organisations from the different sectors and the appearance of knowledge-based innovation systems. URL citations have been defined as the "mentions of a URL in a Web page, whether hyperlinked or not" (Kousha & Thelwall, 2005, p.67), and may appear in the results of a link analysis investigation according to the data collection technique.

Caution is advised against taking the analogy between citations and web links too far. Hyperlinks are placed for a far wider range of reasons than citations, many of which may be considered superficial (van Raan, 2001), they are ephemeral in nature, and unlike citations they have the potential of being bi-directional (although citations may be bi-directional where works-in-progress have been circulated before publication, this is not the norm). The superficial and ephemeral nature of the web means that conclusions drawn from it need to be used with caution; web indicators should be considered 'weak benchmarking indicators' (Thelwall, 2004c), that whilst non-robust can provide formative and semi-evaluative assessment. At the same time it is the superficial and ephemeral nature of the web that provides its richness as an information source, providing the opportunity to view a far wider range of informal as well as formal relationships.

There have been many investigations into what web links may be taken to represent, both at a micro and a macro level. At a micro level there have been a number of classifications of web links to determine the reasons why they have been placed (e.g., Kim, 2000; Smith, 2003; Thelwall, Harries & Wilkinson, 2003; Thelwall & Harries, 2004a; Bar-Ilan, 2004a; 2005b). Whilst at a macro level correlations with recognised indicators of success have been investigated in a number of studies (e.g., Thomas & Willet, 2000; Vaughan, 2004a). There is a need, however, to distinguish between conclusions that may be drawn at a micro level and at a macro level; this has a precedent in citation analysis where critics have often inappropriately criticised the validity of macro studies due to the uncertainty at the micro level (e.g. MacRoberts & MacRoberts, 1989).

Most of the investigations up to now have focused on the interlinking within the academic community, which is unsurprising due to the similarities between webometrics and scientometrics (Thelwall, 2004a), the relative maturity of the academic web (Thelwall, 2001d), the personal interest of academics in the academic community, and the additional insights they can bring to such investigations. Whilst there have been a limited number of investigations into the commercial sector (e.g., Shaw, 2001; Thelwall, 2001b; Vaughan, 2004a; Vaughan & Wu, 2004) and the government sector (e.g., Petricek, Escher, Cox &

Margetts, 2006), there has not been a large scale link analysis of the interlinking between the different sectors on the same scale as those that have looked at the academic community (e.g., Li, 2005). Those studies that have investigated inter-sector interlinking in an attempt to provide more information about knowledge-based innovation systems have either looked at the macro scale with little or no investigation into what an individual connection may represent (e.g., Boudourides et al., 1999; Leydesdorff & Curran, 2000; Heimeriks, Hörlesberger & Van den Besselaar, 2003) or have focused on a very limited study (e.g., Vasileiadou & Van den Besselaar, 2006).

1.6 A link analysis of the United Kingdom

The extent to which a link analysis of the web can provide a new source of information about knowledge-based innovation systems will vary between different countries as they utilise the web to varying extents and in different ways (Thelwall, Tang & Price, 2003). It also depends on a country's representation within traditional bibliometric databases. There is a wide diversity in the proportion of a country's journal coverage in citation databases, such as the Web of Science, and not simply in favour of those English speaking countries (Moed, 2005). The usefulness of patents will also vary from country to country as where there is little patent enforcement, applying for patents are likely to be costly with few benefits. It is therefore appropriate for this investigation to focus on the web within a single country, a natural unit for the investigation of knowledge-based innovation systems (Lundvall, 1992).

There are a number of reasons that make the United Kingdom a natural choice for an extensive link analysis, which is reflected in the large number of link analysis investigations that have already focused on the UK's academic community (e.g., Thomas & Willet, 2000; Thelwall, 2002a; 2003a): the web is well-established within the UK, heavily used by all sections of society; the UK's country-code top-level domain name is heavily used and makes use of an extensive second-level domain name system, e.g., .co.uk. In addition the UK government has recently announced that the next Research Assessment Exercise, the means by which the UK distributes billions of pounds in research funding, is to be the last; replaced instead by an assessment system based on a variety of metrics (HEFCE, 2006). A greater understanding of the role of higher education institutions, and other research based organisations, in the innovation process would help the funds to be distributed in the most effective way.

1.7 Aims and objectives

The aim of this thesis is to determine whether link analysis of the web can provide a new source of information about knowledge-based innovation systems in the UK. The main objectives are as follows:

- Develop an appropriate data collection methodology.
- Investigate what can be inferred from web links about inter-organisational relationships.
- Explore the extent to which inter-organisational relationship information inferred from web links is different to that obtainable by other means.

1.7.1 Developing an appropriate data collection methodology

Until now the majority of webometric investigations have focused on the interlinkages between higher education institutions and what the interlinkages represent. These investigations involve a relatively small number of organisations, with large web sites, that have previously been found to be highly connected. Investigating the web manifestations of knowledge-based innovation systems needs to take into consideration many more organisations: organisations of various sizes and with different web presences. Different types of web sites and new data collection tools becoming available mean that it is necessary to re-examine the suitability of the different methods for data collection that have been used in the past.

1.7.2 Determine what can be inferred from web links

The initial intention was to investigate what can be inferred about inter-organisational relationships from web links through broadly following Thelwall's (2004a) link analysis methodology. Thelwall proposes providing evidence of validity for inferences based on web links through combing a classification of web links using content analysis techniques with correlation tests between a web site's web links and an external indicator of the attribute under investigation. It is necessary to combine classification exercises with correlation tests as correlation tests alone would not show a direct relationship between the web links and innovation, whilst a classification of web links is likely to find that web links are placed for a variety of reasons in addition to those that directly reflect contributions to the innovation process.

1.7.3 Explore the extent that web link derived information is new

The importance of this investigation lies in the potential of the web to provide new insights into knowledge-based innovation systems rather than a repetition of information available elsewhere. It is important to determine whether or not the information is indeed new, or whether the relationships could have been found within traditional bibliometric sources where the information is well organised in comparison to the relative anarchy of the web.

1.8 Research contributions

This study seeks to demonstrate that:

- Data may be collected ethically to provide information about the interconnections between web sites of various different sizes and from within different sectors of society.
- There are significant differences in the linking practices of web sites within different sectors.
- Reciprocal links provide a better indication of collaboration than uni-directional web links, although primarily with affiliated sections of the same organisation.
- The web provides new information about the relationships between organisations, rather than just a repetition of the same information from an alternative source.

1.9 Dissertation structure

In addition to this introductory chapter the thesis is comprised of four main sections:

- The literature review (chapter 2).
- The preliminary investigations (chapter 3).
- The principal investigation: methodology, results and discussion (chapters 4, 5, 6).
- Conclusions of the investigation into the web manifestations of knowledge-based innovation systems (chapter 7).

1.9.1 The literature review

The literature review starts with a brief overview of the terminology used within this thesis, much of which is still in a state of flux within this relatively new field of investigation. After a brief appraisal of alternative approaches to investigations of the manifestations of knowledge-based innovation systems on the web that have previously been undertaken, the literature review is broadly structured according to Thelwall's (2004a) information science approach to link analysis: identification of relevant web pages; data collection; data cleansing techniques; validation through classification of links; and validation through correlation. Although there has not been a large scale link analysis of the interlinking between organisations from different sectors, there have still been a large number of pertinent webometric investigations.

1.9.2 The preliminary studies

The principal investigation is based on the findings of five previously published preliminary studies that have provided insights into what may be inferred from the web about knowledge-based innovation systems and the most appropriate data collection methods.

- *Web crawling ethics revisited: Cost, privacy and denial of service* (Thelwall & Stuart, 2006). Despite the widespread use of web crawlers within webometric research there has been little discussion about the ethical implications of such data collection techniques. This study looks at the ethical implications of their use in webometric research, a necessary step in determining a suitable data collection methodology.
- *What can university-to-government web links reveal about university-government collaboration?* (Stuart & Thelwall, 2005) and *University outlinks: What do links to different domains represent?* (Stuart, Thelwall & Harries, 2007). Utilising the link data collected by the bespoke social science web crawler SocSciBot, two classification studies were carried out to determine the reasons links were placed within the academic domains, and what they could tell us about the relationship between universities and other organisations.
- *Investigating Triple Helix relationships using URL citations: A case study of the UK West Midlands automobile industry* (Stuart & Thelwall, 2006). In line with Thelwall's (2004a) link analysis methodology a small scale pilot study was carried out to determine whether there were enough URL citations between a finite number of web sites from different sectors using the Google's Application Programming Interface, which allows multiple queries to be automatically sent to Google's search engine database in a regulated manner.

- *University-industry-government relationships manifested through MSN reciprocal links* (Stuart & Thelwall, 2007). When the tools provided by search engines for collecting data about the web change it is necessary to reassess the appropriateness of different data collection methodologies, and the potential investigations that may be carried out using the tools. Shortly after the previous paper reporting the results of the pilot study into the West Midlands automobile industry was published, the Live Search search engine introduced a new operator allowing for the collection of more link data. This meant the suitability of utilising web links to investigating knowledge-based innovation systems had to be reassessed.

1.9.3 The main research: methodology, results and discussion

Based on the results of the preliminary investigations and the review of the literature, a final study was carried out in the form of a link analysis of the pharmaceutical industry in the UK, with the results compared to those obtainable through traditional bibliometric sources. Chapter 4 reflects on the reasoning behind a link analysis of the pharmaceutical industry, provides some testable hypotheses, and describes the methods adopted to test these hypotheses. The results of the study are presented in chapter 5, and these are discussed at length in chapter 6.

1.9.4 Conclusions of the investigation into web manifestations of knowledge-based innovation systems

Chapter 7 draws conclusions about web manifestations of knowledge-based innovation systems and the ability of link analysis to provide information on the interactions within knowledge-based innovation systems based on the review of the literature, the preliminary studies, and the large scale link analysis of the pharmaceutical industry within the UK.

2 Review of the literature

2.1 Introduction

Despite the lack of a large detailed investigation into the interlinkages between organisations from the different sections of society using the link analysis approach discussed by Thelwall (2004a), there have been a number of other related investigations: macro studies of knowledge-based innovation systems; investigations into the appearance of co-authored papers on the web; and extensive link analysis of the academic sector.

Within this thesis the term *macro studies of knowledge-based innovation systems* is applied to those studies that have investigated the ability of the web to provide indicators of the interconnectedness of organisations, but have failed to investigate the reasons why links have been placed between web sites at the micro level. Instead they may have looked at the linking structure either on its own (e.g., Boudourides et al., 1999; Leydesdorff & Curran, 2000) or as one layer of a multi-layered communication system (e.g., Leydesdorff, 2001; 2003; Heimeriks et al., 2003) without a fine grained investigation.

The linking structure of the web is not the only potential source of information about the different interlinkages between actors available, although the unique identification provided by URLs make them a popular target for webometric investigations. Citations also provide a unique identifier for specific documents, and their formal structure with identifiable elements has also made them an area of webometric investigation (e.g., Kretschmer & Aguillo, 2004).

The last section of this literature review, and the bulk of the chapter, provides a review of the additional literature relevant to a link analysis approach to investigating the manifestations of knowledge-based innovation systems. In Thelwall's (2004a) proposed link analysis methodology there are a number of distinct stages after selecting the research question and carrying out a pilot investigation: identifying appropriate web pages to answer the research question; collecting the link data; cleaning the data; partially validating the data through correlation; and partially validating the data through a classification of links.

After a brief discussion on the linking terminology used throughout this thesis, this chapter reviews each of these areas whilst, where appropriate, discussing the ideas and concepts with the broader field of bibliometrics and the traditional bibliographic surrogates.

2.2 Key link terminology

Terminology plays an important role in the communication of research if we are to be understood by other researchers and science policy makers (Lazarev, 1996). It is therefore important in such a relatively new field to define the terminology that is used so that it can be used with consistency, and without the confusion that may accompany preconceived notions.

As well as multiple terms being proposed for the application of informetric methods to the web (see section 1.4) there has also been a variety of terminology used to describe the links between different web pages. For example, inlink (Björneborn, 2004), backlink (Harter & Ford, 2000), hypertext citation (Chen, Newman, Newman & Rada, 1998) and sitation (Rousseau, 1997) have all been used to describe the link one web page has pointing to it from another within webometric investigations. This thesis employs the *inlink* and *outlink* terminology that Björneborn (2004) utilised in his attempt to create a consistent terminology

and is now widely used within the webometric community, although this is by no means the only terminology currently being used (e.g., Chau, Shiu, Chan & Chen, 2007).

Before discussing the terminology of the links between web documents any further it is necessary to briefly discuss what is meant by a web document. Prior to the introduction of the concept of Alternative Document Models (ADMs) by Thelwall (2002d), the web page formed the basis of link analysis investigations, with the links to and from individual web pages being analysed. Thelwall argued that the web page was not necessarily the most appropriate unit for link analysis, but rather it may be more appropriate to aggregate a number of web pages into a single document, suggesting the directory, the domain, and multiple domains as ADMs within the academic arena. Whilst Thelwall has focused primarily on aggregating through lexical components of URLs (Thelwall & Price, forthcoming), this is not the only possibility; web pages may also be clustered into web documents, or 'web sites', semantically, i.e., according to the similarity of their content; or topologically, i.e., according to how the web pages are interlinked (Cothey, Aguillo & Arroyo, 2006). Lexical aggregation of web pages may be considered the most popular due to the relative simplicity.

The choice of web document provides context to the link terminology: inlink, outlink, reciprocal-link, and self-link:

- *Inlink:* A link to a web document from a web page not included within the web document.
- *Outlink:* A link from a web document to a web page outside the web document.
- *Reciprocal-link:* A link from one web document to another web document which, in turn, has a web link back to the original web document.
- *Self-link:* A link from a web page in a web document to the same or another web page in the same web document.

These definitions differ slightly from those of Thelwall (2004a), who makes the presumption of the web page as the web document unless otherwise stated. As alternative models exist, and have been found to be more appropriate in certain situations (Thelwall, 2002d), the notion of a default web document seems passé, and the application of the web document as a qualifier is a necessity.

It is also useful to briefly discuss the terminology that is applied to the components of URLs. URLs can consist of a number of different parts: a protocol, a user, a password, a domain name, a port, a path, a query, and an anchor. From a link analysis perspective it is necessary to clearly define the terminology that is utilised for discussing the domain and the path sections of the URL, as these are the parts that are used most often within the lexical aggregation, and are further subdivided. They are also the parts for which there is conflicting terminology. The terminology used throughout this thesis to refer to parts of a URL is illustrated in Figure 2-1 with the example of a typical static URL.

The top-level domain name (TLD) may consist of either a country-code top-level domain (ccTLD), e.g., .uk or .fr, or a general top-level domain (gTLD), e.g., .com or .org. Many of the ccTLDs have been further sub-divided with the use of second-level domain names (SLDs) for different types of organisations. For example the .uk ccTLD has been further divided through the use of 13 SLDs some of which have restricted use, e.g., .nhs.uk, and others which, whilst aimed at a particular type of organisation, are openly available to all, e.g., .org.uk. Not all ccTLDs are further divided, e.g., .fr and .dk do not incorporate SLDs, and

even where there are SLDs there may be some sites which fall outside the regular domain structure. For example, the British Library's web site, www.bl.uk, predates the current domain name system.

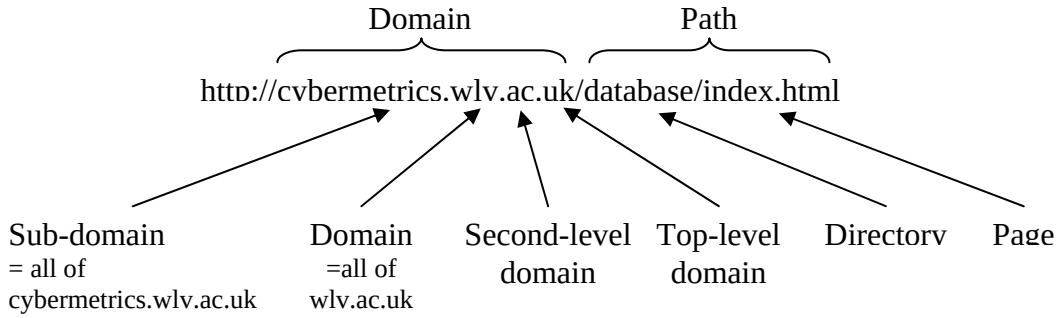

Figure 2-1 An example of a static URL labelled with the link analysis terminology

2.3 Macro studies of knowledge-based innovation systems

Ingwersen and Björneborn (2004) distinguish between three levels of link analysis investigation: the macro level, the meso level, and the micro level. They define macro analyses as those investigations that study the clusters of many web sites or TLDs, meso analyses as those that study larger sub-sites and sites, and micro analyses as those that investigate the web pages, web directories, and small 'sub-sub-sites'. The terms are not being used to describe different methodological approaches, but rather to refer to different aggregates of web pages to form web documents; the terms could equally be applied to the aggregating of web links in ADMs. In this study the terms macro and micro analyses are applied to distinguish between studies that have looked at the linking between web documents without any analysis of the individual links (macro analyses), and those that focus on what individual links represent (micro analyses), following Bar-Ilan's (2004a) use of the term microscopic link analysis.

There have been a number of studies that have investigated the potential of the web to provide indicators of knowledge-based innovation systems, either on its own (e.g., Boudourides et al., 1999; Leydesdorff & Curran, 2000) or in association with other layers of the communication network (e.g., Leydesdorff, 2001; 2003; Heimeriks et al., 2003). These studies have primarily focused on the Triple Helix vision of the knowledge-based innovation system first put forward by Etzkowitz and Leydesdorff (1995), which sees the Triple Helix interactions of academia, industry and government as the key component to any innovation strategy, with academia having an increasingly important role in the knowledge economy (Etzkowitz & Leydesdorff, 2000). One criticism of the Triple Helix model has been that it is a highly abstract and un-testable systems theory (O'Malley, McOuat & Doolittle, 2002). Whilst this may be the case, it is the Triple Helix's abstract nature with a lack of reified system boundaries (Etzkowitz & Leydesdorff, 2000) that makes it a popular model for investigation. The system may be defined for each project in turn, providing a useful empirical tool for examining the different interactions (Leydesdorff & Etzkowitz, 2003). Whilst the original Triple-Helix focused on the interactions between academia, industry, and government, not even this should be reified, rather alternative sectors may be substituted depending on what is under investigation. For example, the triple helix of the public, academia and government has

been suggested to represent the dynamic of the scientific controversies (Etzkowitz & Zhou, 2006).

The first reported investigation of the manifestations of knowledge-based innovation systems on the web was Boudourides et al.'s (1999) draft report into what they referred to as the 'triple-helix-ness' of the web space within Europe and the US, defining 'triple-helix-ness' as the interrelationship and connectivity between universities, governments and industry. Using link data collected with the then popular AltaVista search engine they studied the triple-helix-ness of 112 web sites by analysing certain of the web sites' characteristics:

- Whether they linked to the government sector.
- Whether they linked to the university sector.
- Whether they linked to the industry sector.
- The number of incoming links.
- The type of institution to which the web site belonged.

Using multiple correspondence analysis it was concluded that within the US the government sector is by far the most triple-helixed, whilst within Europe the university sector was the most triple-helixed. Whilst the study is important for its recognition of the potential of the web to provide information about an organisation's relationship with organisations from other sections of the Triple Helix, as a draft study there are many limitations. There are no reasons given for the selection of the web sites included in the study, and it makes no attempt to either define the university, industry and government spaces it is looking for outlinks to, or whether such web spaces should be considered to be different for Europe and the US. The investigation also fails to take into consideration differences in the nature of the links or indeed the number of links from a web site; an organisation that has many links reflecting collaborative relationships with numerous organisations from each of the sectors is only considered as triple-helixed in their investigation as an organisation with a single link for information purposes to each of the sectors.

Following the lead of Boudourides et al. (1999), Leydesdorff and Curran (2000) also used the advanced search technology of AltaVista to investigate the similarities and differences between national systems of innovation by utilising the ccTLDs. Their investigation compared the Triple Helix of university, industry and government relations that appeared within the Dutch domain (.nl), the Brazilian domain (.br) and the gTLDs (e.g., .com, .org, .net). The investigation analysed the appearance of the terms 'university', 'industry' and 'government', and combinations thereof, within the different domains. It also investigated the appearance of those terms within the anchor text of the links from web sites within the different domains. Leydesdorff and Curran (2000) concluded that the web was suitable for investigations into Triple Helix relationships, although this assessment was made primarily on the key word searches as they found the appearance of the key words in links to provide very few items. As with Boudourides et al.'s (1999) earlier investigation, the study made no attempts to validate the data they retrieved as providing information about any particular type of relationship, or whether the appearance of a term within a link meant that the web site being linked to was part of that sector.

Priego (2003) attempted to determine the Triple Helix nature of two research centres through an analysis of the research centres' outlinks using a vector space model for similarity assessment. Unlike the previous investigations of Boudourides et al. (1999) and Leydesdorff and Curran (2000), the outlinks were collected through a web crawler rather than a search

engine, and the sectors that the linked-to web sites were part of were determined by a classification exercise rather than only looking for those that were within a certain TLD or which had a keyword within the outlink text. However, once again, there was no attempt to determine the types of relationship that were reflected by the web links.

The lack of any type of link classification may be justified within each of the studies on the basis that although the exact nature of these connections is unknown there is nonetheless an identifiable connection between the different sectors. It is important, however, to be able to understand the nature of these relationships, especially when comparing heterogeneous sections of the web. Recognition of the need for greater understanding of what is being shown by the web links has been reflected in those studies that have looked at the web manifestations in conjunction with other 'layers' of the communication network: bibliographic and patent databases (Leydesdorff, 2001; 2003); academic journals and project collaborations (Heimeriks et al., 2003).

Leydesdorff (2001) investigated the appearance of the trade names and the generic names for different drugs on the web and within traditional bibliographic databases: the European patent office, the Medline database, the Science Citation Index, and the Derwent Patent Citation Index. The gTLDs on the web were investigated through the use of the advanced search operators of the AltaVista search engine. As with his earlier investigation (Leydesdorff & Curran, 2000), rather than understanding the relationships between any particular actors, the investigation focuses on the creation of indicators for attributes that are found within the knowledge-based innovation systems, in this case the appearance of the trade names and generic names of drugs. Whilst Leydesdorff concludes in this study that the internet is "so overwhelmingly commercial that it is no longer useful as an indicator of 'user' interests", it would probably be more accurate to say that there is a need for greater control of the selection of web sites under investigation and improved data cleaning techniques if useful indicators of user interests are to be established.

In Leydesdorff's (2003) later paper he focuses more on the changing nature of the Triple Helix over time, as manifested by changes in the appearance of the terms university, industry, and government, and combinations thereof, within the web generally, specific domains on the web, the Science Citation Index, and U.S. patent data. Drawing conclusions about the nature of the institutional relationships between industry, government and academia, based on the appearance of the three terms is tentative at best; however, inclusion of the time variable and the application of the theory to the web are beyond the capabilities of the AltaVista search engine. That a search engine allows the incorporation of a time restrictive operator is not the same as the operator providing robust results. The nineties saw huge changes in the types of organisation that utilised the web, old web pages have disappeared, or been written over, and the search engine is likely to have changed its searching policy numerous times as the web has grown and their computing power has grown.

Heimeriks et al.'s (2003) investigation compared formal scholarly communication in academic journals, the communication network exhibited in project collaborations, and the communication network shown on the web, in an attempt to clarify the different functions of the different networks. Unfortunately the networks were biased towards the academic community and formal collaborations, and as such the full potential of the web as an information source on less formal collaborations between wide ranges of organisations was not utilised. The organisations that were selected as nodes for the journal network were those within the European Union that had published in a journal indexed by the Science Citation Index (SCI), which are predominantly academic institutions; this academic bias continued into

the collaboration network, as the institutions identified in SCI formed the basis of the collaboration network; and the academic and formal collaboration bias continued into the web network as it was between these organisations that web links were looked for. The investigation did not allow for organisations to appear in the web network unless they had already appeared in one of the earlier networks. The web network was a very weak network, but this may be attributable to the data collection technique; whilst a web crawler was used, it only crawled two levels deep. Although Heimeriks et al. draw the conclusion that the web is used merely for communications with users of the knowledge resources, rather than being related to the co-production of knowledge and collaboration, a larger, more detailed, investigation is necessary before such a statement can be made with any strength.

The Triple Helix and the other knowledge-based innovation systems provide useful frameworks for analysing the relationships between the different sectors of society, however it is important that Leydesdorff's (1987) cautionary note regarding the measurement of science through bibliographic surrogates is heeded by webometricians as well: "scientometricians…have often been too eager to produce meaningful results based on what they happen to be able to measure in science" (p.305). It is necessary that we continue to look at the micro level in conjunction with the macro level if meaningful conclusions are to be drawn from the data.

2.4 Other web manifestations of organisational interlinkages

As more academic work appears on the web it is unsurprising to find growing interest in the appearance of web citations, both as an indicator of the impact of a journal article (e.g., Vaughan & Shaw, 2003; 2005), and as an indicator of collaboration. Within traditional bibliometric investigations co-authorship has been widely used as a proxy of collaboration (Bordons & Gómez, 2003), with the author and organisational affiliations providing an opportunity for collaboration to be investigated at various levels of aggregation: individuals, institutions, and nations (Glänzel & Schubert, 2004). Whilst co-authorship collaboration will not always reflect the same types of relationships between the authors and their respective institutions (Katz & Martin, 1997), it must be viewed as a strong indicator of operationalized collaboration.

Kretschmer and Aguillo (2004) in a comparison of co-authorship patterns in traditional bibliometric databases and the network visible on the web found that a high proportion (78%) of multi-authored publications were visible through a search engine, even though no links were found between the home pages of any of the authors in the study. These findings were reiterated in a later investigation where it was concluded that "counts of hyperlinks are not useful in reflecting the collaboration structures as measured by bibliographic data" (Kretschmer, Kretschmer & Kretschmer, 2007, p. 536). Although more web links may have been found if an alternative web document had been used rather than a single web page, it does not detract from the finding that there is a lot of potentially useful information on the web that does not necessarily appear in the web's link structure. Although Kretschmer and Aguillo (2004) started with a list of articles, it is not necessarily essential to have a list of published documents to begin with. For example, Bar-Ilan (2000b) investigated the bibliographic references found in the web pages returned for the search engine query 'informetrics or informetric', finding that the results were often better than those of the more expensive bibliographic databases.

Automatic extraction of references from web pages seems a possibility, as they are meant to conform to certain conventions. Nevertheless, such information would still only provide information about the formal, mainly academic, collaboration that is already available in bibliographic databases. There are also likely to be increased difficulties in distinguishing between homonyms as the identified documents on the web do not necessarily have the affiliated institution details that are included within bibliographic databases.

2.5 Link analysis

Although there has not been a large scale link analysis of the interlinking between organisations from different sectors, there have been a large number of webometric investigations that are pertinent to this issue. This section of the chapter reviews the additional relevant literature using the framework of Thelwall's (2004a) link analysis methodology:

- Identifying web pages appropriate to the research question.
- Data collection.
- Data cleaning.
- Partially validating link count results through correlation tests.
- Partially validating the interpretation of the results through a link classification exercise.

The framework of Thelwall's (2004a) link analysis methodology is a more appropriate structure than discussing each of the studies individually as it is their constituent parts that are most important to this investigation.

2.5.1 Identifying web pages relevant to the research question

Identification of the web pages relevant for a particular investigation is often the part of a link analysis investigation that is brushed over in an attempt to get on with the 'proper' investigation. Nevertheless, the selection of web pages, or group of web pages, for inclusion within a link analysis investigation needs to be approached methodically as there are few clear cut boundaries on the web. There are two stages for the identifying of relevant web pages, the conscious decision on the part of the investigator about which organisations they want to include, and then a decision about how they are going to operationalize the first decision. The focus within this section is necessarily on the stage where the conscious decisions are operationalized. In practice the stages are not necessarily linear or clear-cut, but rather are a constant to-and-fro between the ideal and the feasible.

The structure of URLs enables a choice to be made for the inclusion of whole sets of web sites with particular gTLD name (e.g., .com, .org, .gov), specific ccTLD names (e.g., .uk, .fr, .se), or specific web pages. Since certain countries have further divided their ccTLD into second level domain names (e.g., .co.uk, .org.uk, .nhs.uk), these provide both an indication of the type of web site as well as its geographic location. Utilising these gTLDs or SLD/ccTLDs provides a simple way of looking at the connections between different sectors of society (e.g., Thelwall, 2001c), different countries (e.g., Thelwall, Tang & Price, 2003), as well as the links between specific web sites and gTLDs or SLD/ccTLDs (e.g., Li, Thelwall, Musgrove, & Wilkinson, 2003). Whilst the choice of a gTLD or SLD/ccTLD is generally made on an intuitive basis, there are numerous caveats that should be taken into consideration before drawing conclusions: ccTLDs are often overlooked by organisations choosing a domain name

in favour of the more popular gTLDs; ccTLDs may allow foreign registration and may be used to create novelty URLs, e.g., del.icio.us, or to reflect the type of organisation, e.g., the .tv ccTLD is now more associated with television industry than the island nation of Tuvalu. Recall also that there are important web sites that predate existing domain name structures e.g., the British Library web site www.bl.uk predates the use of second level domain names within the UK; and most importantly, whilst many of the gTLDs and second level domain names are meant to reflect a particular type of organisation, few are regulated.

Identification of specific web pages or web sites, as opposed to gTLDs or ccTLDs, can be achieved in a number of ways: through the selection of random web sites; searching for known organisations on the web; searching for unknown organisations on the web; or through the identification of web pages by experts.

Selection of random web pages has been achieved in three ways: through the use random web page generators that have been provided by some search engines (e.g., Haas & Grams, 1998; Koehler, 1999); through randomly generating Internet Protocol (IP) addresses (e.g., Thelwall, 2001b); and through the random selection of web pages from a web crawl (e.g., Thelwall, Harries, et al., 2003). Whilst a search engine's random URL generator would seem to offer a quick and easy method of identifying random URLs, it is necessary that all such web tools are used with caution. Random URL generators have a number of limitations: some have been found not to be random (Koehler, 1999); the search engines they get their results from are biased (Thelwall, 2001b) (see section 2.5.2.3); and there are no details of how the random page is picked, for example, whether all web sites are weighted equally or all pages are weighted equally. Whilst the selection of random IP addresses provides an alternative method of randomly selecting web sites, Thelwall (2001b) acknowledges that there are still limitations in as much as multiple domain names may be found on a single server (which is now probably a fatal limitation), and conversely a single domain name may use multiple servers. In addition, searching for random IP addresses will result in numerous 'File not found' error messages. Whilst the ratio of servers found to servers not found for the randomly produced IP address may have been efficient in 2000 when Thelwall's investigation was carried out, there is a need to reassess the suitability of such methods with the introduction of Internet Protocol version six which contains many more possible IP addresses (2^128) than the previous version four (2^32) (Lawrence & Giles, 1999), and which will run in conjunction with IP version four for the foreseeable future. Whilst randomly selecting web pages from the data collected through a web crawl allows the researcher to control the way a page is randomly selected, it is restricted by the extent of the web crawl (see section 2.5.2.2).

The principal method of identifying relevant web sites for inclusion in a webometric investigation is through identifying the web sites for a predetermined set of organisations. Whilst there are many examples within the academic community where there are a recognisable set of higher education institutions (e.g., Chen et al., 1998; Smith, 1999a; Thomas & Willet, 2000), there are also predefined sets of commercial organisations that have been used within link analysis investigations (e.g., Vaughan, 2004a; Vaughan & Wu, 2004). Identification of the relevant web sites for specific organisations is accomplished either through the use of a search engine or through a page of links to each of the organisations from a recognised source. However, there is often more than one web site for an organisation. Most of the studies have operationalized the concept of the web site lexically, utilising either the main web site only, or combining the results with other recognised domain names. Thelwall (2002b) describes identification of all the non-derivative domain names used by British universities as impossible, a statement that must equally apply to global commercial

organisations that are likely to have multiple web sites in multiple countries for different purposes. Conclusions drawn on such studies can be severely affected by the choice of site structure by an organisation without closer examination.

Without an initial set of organisations for which to find web sites, search engines and web directories can also provide a way of finding web sites on a particular topic (e.g., Larson, 1996; Prime, Bassecoulard & Zitt, 2002), whilst some sites provide lists of the most popular web sites that can be used (e.g., Shaw, 2001). It is necessary, however, to take into consideration the biases of search engines (see section 2.5.2.3), how internet-use statistics are being calculated, and to recognise that many important sites may be missed because they are either not indexed by the search engine or they do not contain the particular term that was queried. Searching for web sites that fit a particular topic also fails to recognise whether there are certain important organisations to the investigation that may not have a web site.

It is also possible to have appropriate web pages identified by an expert within the relevant field. For example, recognising the popularity of pornography on the internet and the lack of porn sites provided by a list of the most popular sites, Shaw (2001) added two 'expert-identified' porn web sites to her investigation. Whilst a subject expert may be of use in checking the comprehensiveness of a list of web sites, and possibly proposing additional important web sites that have not been included, there are obvious limitations in relying on a single person to provide a comprehensive list of appropriate sites for anything but the most specific of subjects.

The size of the web and the lack of standardisation in the way that it is used mean that there is a lack of clear cut boundaries to a link analysis investigation. Even the relatively simple process of identifying the web pages appropriate to an investigation needs to be couched with caveats.

2.5.2 Data collection

Link data may be collected in three ways, either through manual browsing (e.g., Kretschmer & Aguillo, 2004), through the use of a personal web crawler (e.g., Chen et al., 1998; Terveen & Hill, 1998; Thelwall, 2001a), or through the manipulation of the data collected by search engines (e.g., Rousseau, 1997; Vaughan 2004a). Selection of one data collection method over another may be viewed as a trade-off between the comprehensiveness of coverage and the practicalities of data collection. Whilst a search engine provides access to a greater amount of information than an individual researcher could hope to collect on their own, it still does not give access to all the available information and the exact nature of a search engine's collection policy is not revealed.

It is not unusual for a link analysis investigation to take advantage of the different benefits provided by the different data collection methods by using one technique in one part of an investigation, and an alternative technique in another part. For example, Li et al. (2003) calculate inlinks from the whole web with a search engine, whilst links between a finite set of web sites are investigated through the data collected with a web crawler. As discussed above, even where an alternative to a search engine is used for data collection, a search engine may still play a pivotal role in the investigation through the identification or selection of web sites.

2.5.2.1 Manual data collection

The limitations of manual browsing are fairly obvious when one considers the size of the web, one recent estimate of which put the indexable web at more than 11.5 billion pages (Gulli &

Signorini, 2005), and such an estimate may be considered the tip of the iceberg. Within their study they define the indexable web as those pages that are considered for indexing by one of the major search engines. Many pages are not indexed by the search engines either due to policy decisions or technical limitations on the part of the search engines (Bar-Ilan, 2004c). The time consuming nature of manual browsing for link data, and the potential for human error when links are not necessarily highlighted and may be hidden behind pictures, means that for all but the smallest of investigations it is only recommended for checking the completeness of other methodologies (Almind & Ingwersen, 1997).

For a relatively small number of web pages, for example an investigation into the linking between 17 academic homepages (Kretschmer & Aguillo, 2004), manual browsing allows the investigator to cut down on the element of uncertainty involved in the use of web crawlers and search engines. For example, whereas a search engine may leave a user in doubt as to whether a particular page has been indexed or whether a web crawler has recognised all the links embedded within a web page, theoretically an investigator can be sure of their own meticulous investigation.

2.5.2.2 Personal web crawlers for data collection

Web crawlers, also known as spiders or robots, are programs that download web pages automatically, extract the embedded URLs, and then fetch them iteratively (Thelwall, 2001a). Theoretically a crawler can be instructed to crawl as little or as much of the web as required by the researcher. In practice, however, crawlers have a number of restrictions: the linking structure of the web; the capabilities of the crawler; and the computing power and bandwidth available. Although the term ‘web robot’ has also been applied to a program that automatically submits searches to a search engine (Larson, 1996) within this thesis such programs are referred to as ‘search engine scrapers’ to differentiate between the two.

The amount of the web that can be retrieved by a web crawler depends very much on the seed list of URLs that are initially given to the web crawler for it to download and extract the embedded links from. This is best illustrated by Broder et al.’s (2000) ‘bow tie’ model of web connectivity, based upon two AltaVista web crawls of over 200 million web pages and 1.5 billion web links (see Figure 2-2). Web pages can be classified as being in one of a number of distinct components according to how, through their links with other web pages, they are linked to the web as a whole. The Strongly Connected Component (SCC) consists of the core of the web, a highly connected set of pages from which any web page in the SCC can be reached from any other web page in the SCC through following a path of links. The IN component consists of pages that lead to the SCC but cannot be reached from it, whilst the OUT component consists of pages that can be reached from the SCC but do not lead to the SCC. In addition there are pages that are connected to the main network of the web, but are not connected either to or from the SCC, the tendrils and tubes in the diagram. The disconnected components are those pages that are in no way connected to the SCC. If the seed list of web sites are all taken from the disconnected components, or the ‘out’ section of the model there is no way for the crawler to find those pages that fall within the ‘in’ part of the web or the strongly connected component of the web. It is quite likely that a single organisational web site might have pages that fall into different sections of the model, including the disconnected components, and as such there will be pages that cannot be found unless there are multiple specific seed URLs from which the crawl is started.

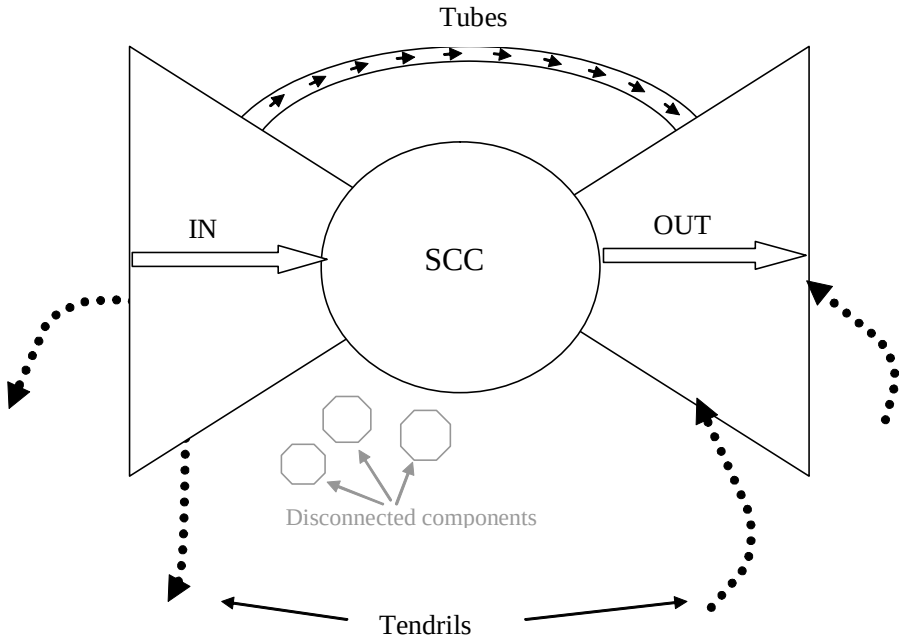

Figure 2-2 Broder et al.'s (2000) web connectivity model

There are various different web crawlers available, both academic and commercial, content crawling and URL crawling. Each of the crawlers provides the researcher with different options as to the parameters that can be set and the type of information that can be collected. As Arroyo (2004) found in a comparison of seven commercial crawlers and an academic crawler, and had been found in Koehler's (1999) earlier study using two different web crawlers, the choice of crawler can have repercussions in the number of pages that are found within a web site, especially with regards to the finding of dynamic pages. There have been a large number of webometric investigations using both the commercially available crawlers (e.g., Haas & Grams, 1998; Terveen & Hill, 1998; Koehler, 1999; Heimeriks et al., 2003; Priego, 2003), as well as the bespoke crawlers designed for webometric investigations (e.g., Thelwall, 2001c; 2001e; 2002d; 2002e; 2002h; 2003b; 2003c; 2003d; 2004b; 2005a; Thelwall & Harries, 2004b; Thelwall, Harries, et al., 2003; Thelwall & Price, 2003; Thelwall & Wilkinson, 2003b, 2004; Wilkinson, Harries et al., 2003; Ajiferuke & Wolfram, 2004; Harries, Wilkinson, Price, Fairclough, & Thelwall, 2004; Vasileiadou & van den Besselaar, 2006) such as the SocSciBot web crawler (Thelwall, 2002f; 2003a) and other freely available academic crawlers (Chen et al., 1998). The advantage of the bespoke web crawlers is that they give greater control in the setting of parameters, and allow for greater understanding in what is being crawled, and why pages may not be crawled. Commercial crawlers are liable to have many of the details regarding the identification of URLs and duplicate pages hidden from the user.

The duplicate page checker is an important aspect of the web crawler, and a reason for potential differences in the number of pages retrieved by different web crawlers for the same web site. Whilst some web crawlers may be described as content crawlers and attempt to prevent duplication of web pages by comparing the content of the page they are downloading with the content of the pages they have already downloaded, others may only check for duplication in the URLs of the web pages (Thelwall, 2004a).

One of the biggest restrictions on the use of web crawlers is caused by the processing power and bandwidth necessary for downloading the millions of web pages that may exist within a single web site, let alone the billions that exist on the whole of the web. As such a web crawler is only suitable for investigating a selection of web sites rather than the whole of the indexable web due to the facilities available to most researchers. There are also issues that need to be addressed regarding the effect crawling has on the servers that are being crawled, the ethical issues regarding using web crawlers in webometric investigations are discussed in greater detail in the first of the preliminary studies in the next chapter (see section 3.2).

Attempts to limit the amount of data that is downloaded using web crawlers within webometric investigations for ethical or other reasons means that often it is not the whole of a web site that is crawled, instead either a specific number of pages are crawled (e.g., Terveen & Hill, 1998) or more usually all the pages up to a specific depth are crawled (e.g., Haas & Grams, 1998; Heimeriks et al., 2003; Vasileiadou & van den Besselaar, 2006). Such restrictions run the risk of invalidating the investigation if the depth is not deep enough. For example, on finding a lack of links between web sites Heimeriks et al. (2003) had to conclude that it was possible that it was the depth of the crawl that had proved to be insufficient rather than a lack of linkages between web sites.

The Statistical Cybermetrics Research Group at the University of Wolverhampton has been running a complete crawl of all the higher education institutions within the UK annually since 2000, as well crawling occasionally other academic institutions from around the world (Statistical Cybermetrics Research Group, 2007), and these databases have formed the basis of many of the link analysis investigations (e.g., Thelwall, 2001a; 2001e; 2002d; 2002e; 2002h; 2003b; 2003c; 2003d; 2004b; 2005a; Thelwall & Harries, 2004b; Thelwall, Harries, et al., 2003; Thelwall & Price, 2003; Thelwall & Wilkinson, 2003b; 2004; Wilkinson, Harries et al., 2003; Harries et al., 2004). However, as has already been mentioned, there is a need to look beyond the academic community in isolation, and large scale investigations of the interlinkages between organisations from different sectors is beyond the limits of web crawls by a small research group.

2.5.2.3 Search engines

The major search engines have their own web crawlers which index far more of the web than a small research group could hope to, and the information they provide has been used for data collection in numerous webometric investigations (e.g., Larson, 1996; Rousseau, 1997; Ingwersen, 1998; Smith 1999a; 1999b; 1999c; 2003; 2004; 2005; Leydesdorff & Curran, 2000; Thomas & Willet, 2000; Thelwall, 2001c; 2002a; 2002b; 2002c; Thelwall & Smith, 2002; Musgrove, Binns, Page-Kennedy, & Thelwall, 2003; Qiu, Chen & Zhi, 2003; Tang & Thelwall, 2003a; 2003b; Thelwall & Tang, 2003; Thelwall & Harries, 2004a; Bar-Ilan, 2004a; 2004b; 2005b; Vaughan 2004a; Vaughan & Thelwall, 2003a; 2003b; Vaughan & Wu, 2004). In addition to general keyword searches they often provide operators that allow for the searching of the text within specific parts of pages, including URLs, and to search for links from the web that point to specific pages or web domains. Used in conjunction with Boolean algebra, webometricians are able to build complicated queries that can retrieve very specific information.

For certain link analysis investigations search engines are the only option due to the amount of the web the researchers want to include. Many investigations want to include as much of the web as possible: investigations into why certain web pages have inlinks and who

from (e.g., Rousseau, 1997; Smith, 2004; Thelwall & Harries, 2004a; Vaughan & Thelwall, 2003a); studies which base metrics on the number of inlinks (e.g., Ingwersen, 1998; Smith, 1999a; 1999b; 1999c; 2003; 2005; Thomas & Willet, 2000; Thelwall, 2002a; Qiu et al., 2003; Tang & Thelwall, 2003a; Vaughan, 2004; Vaughan & Thelwall, 2003b); comparisons of inlinks with self-links (e.g., Bar-Ilan, 2004b); comparisons between large sections of the web (e.g., Leydesdorff & Curran, 2000); or investigations into the interlinking between large sections of the web (e.g., Thelwall, 2001c; Thelwall & Smith, 2002; Musgrove et al., 2003; Tang & Thelwall, 2003b). Search engines may also be used for the retrieval of additional information, for example, a date-stamp they apply when they first index a web page (e.g., Leydesdorff & Curran, 2000) or a web page's language (e.g., Thelwall & Tang, 2003), although there is a need for detailed investigation as to how such attributes are assigned before utilising them within a webometric investigation.

Other investigations have used search engines for their relative ease of use even though the same study would have been possible with a personal web crawler: Larson's (1996) co-link analysis looked at the outlinks of just 34 web sites; Thelwall (2002b, 2002c) looked at the interlinking between 86 UK university web sites, the focus of many later personal web crawler based investigations; and Bar-Ilan (2004b, 2005b) researched the reasons for link placement between eight Israeli universities.

In spite of the ease of use, and the large amount of data that can be collected from search engines, it should be recognised that the priorities of the commercial search engines are not the same as the academics who use them in webometric research (Thelwall, 2000; Bar-Ilan, 2001). Whereas a lack of consistent results, consistent operators, open crawling and ranking policies may do little to detract from the average user's experience of a search engine, such things are a core part of academic investigations and the use of search engine data in academic investigations has been questioned (Snyder & Rosenbaum, 1999).

The inconsistency in the operation of search commands is most noticeable through the addition of two mutually exclusive searches, which between them should include all pages indexed, but do not sum to the total number of pages indexed. Smith (1999a) provides the following example of searches and their number of queries returned:

- link:auckland.ac.nz/ 14796
- link:auckland.ac.nz/ AND host:auckland.ac.nz/ 4629
- link:auckland.ac.nz/ AND NOT host:auckland.ac.nz/ 10616

In the above example the first query should return all the links that point to the auckland.ac.nz web site, which should theoretically be the sum of those web pages that link to the auckland.ac.nz web site and are hosted within the auckland.ac.uk domain (the second query) and those web pages that link to the auckland.ac.nz web site and are not part of the auckland.ac.uk domain (the third query). The same web page can not be both within the auckland.ac.nz domain and without the auckland.ac.nz domain, but must be one or the other. Despite this the addition of the exclusive searches adds up to 15,245, more than the initial search.

There is also inconsistency in the number of pages found by a search engine for the same query. Over a period of time, changes in the numbers of pages found by search engines operators may be partially attributable to changes in the number of pages on the web, although

this is not the only reason for changes in the number of pages found. Numbers are found to fluctuate over a short period of time (Bar-Ilan, 1999; Rousseau, 1999; Mettrop & Nieuwenhusen, 2001), and even if the numbers are the same or similar the actual hits may be different (Bar-Ilan, 1999). It has been suggested that the primary reason for the differences in results is attributable to search engines having multiple databases, each of which may have different crawls of the web, to cope with high volumes of traffic (Thelwall, 2001f), and this has been confirmed for the Google search engine (Google Librarian Central, 2007). Whatever the reasons, it is widely recognised that search engine results are irreproducible (Rousseau, 1997; Bar-Ilan, 1998).

Webometricians also have to endure potential volatility in search engine features. Whilst this is more likely to effect longitudinal studies that investigate the results from search engines over a period of time, there is the possibility of disruption to any webometric investigation. The validity of a number of the early webometric studies can no longer be checked due to the search engine features no longer being available. For example, the WebCrawler random URL generator used by Koehler (1999) was later reported to be inactive (Koehler, 2002), whilst it is not unknown for a feature that has been used in a webometric study to become obsolete before the results have been published (Tang & Thelwall, 2003). During the period of this investigation an overhaul of Google's API service saw a restriction in the support of their SOAP protocol (Google, 2006), whilst MSN's Live Search announced the removal of some of its key link operators (Live Search, 2007). It is not only the features that may disappear, in some cases the search engines themselves have disappeared, or at least the databases behind them have. For example, a longitudinal study of the web pages containing the term 'informetrics' between June 1998 and 2003 had to utilise different search engines as some of the original search engines either were not available or served the results of another search engine (Bar-Ilan & Peritz, 2004).

Although search engines provide access to more data than could be collected through a personal web crawler they still do not crawl the whole of the web, something early webometric studies exhibited a certain naïveté about. For example, Larson (1996) mentions the Inktomi search engine's complete crawl of the web without any qualification, something that is clearly impossible unless they could be sure of having a seed list that included a web page within every one of the disconnected components as well as the end of every tendril linking into the OUT section (see Figure 2-2). In 1998 Lawrence and Giles (1998) stated that there was no search engine indexing more than a third of the indexable web, and based on such findings Bar-Ilan (2001) states in her review of data collection methods on the web for informetric purposes that it "is definitely not enough to use the currently largest search engine". Thelwall's (2000) position is rather less dismissive of the single search engine, rather he suggests that before using a search engine for calculations, checks should be made to ensure that coverage of pages is high.

The subject of what is and what is not indexed by a search engine has societal as well as webometric ramifications, as they are the principal method through which people find information, and the biases regarding what is, and what is not crawled have been the focus of a number of investigations. Vaughan & Thelwall (2004) concluded that biased coverage in favour of U.S. web sites, compared with those of China, Taiwan and Singapore, appeared to be caused by technical and historical reasons, e.g., early adoption of the web in the U.S., whilst other studies have suggested that political and economic motivations may also be involved (Introna & Nissenbaum, 2000; Van Couvering, 2004). Although there are

considerable implications of search engine bias, the notion of deliberate bias for political and economic motivations has not yet been proved.

Recognising the limitations of a search engine's crawling, indexing, and ranking capabilities has led to various evaluations of available search engines. Whilst early evaluations focused on the facilities provided by the then current crop of search engines the fast changing nature of the search engine sector means that useful evaluative papers have to include evaluative criteria and methodologies that may be applied to new search engines (e.g., Chu & Rosenthal, 1996; Schwartz, 1998; Oppenheim, Morris, McKnight & Lowley, 2000; Vaughan, 2004b). Not all the suggested search engine evaluation criteria are suitable for assessing a search engine for link analysis investigations. For instance, response time (Chu & Rosenthal, 1996) would be unlikely to play a decisive role in choosing a search engine for link analysis, unless it was going to mean the investigation was to take an unacceptable period of time, whilst the popular precision and relevance criteria (Chu & Rosenthal, 1996; Oppenheim et al., 2000; Vaughan, 2004), as well as the closely connected recall (Oppenheim et al., 2000; Vaughan, 2004), have less to do with a user's opinion on whether they are relevant or not, but more to do with the accuracy of the search engine to fulfil the search query. It may even be that a search engine's ability to fulfil traditional concepts of relevance may have a negative effect on a webometric investigation. For example, in addition to retrieving pages that contain a keyword, a search engine may retrieve web pages that are linked to by a number of pages containing the relevant keyword, or pages that contain synonyms of the keyword.

Smith (1999b) states that it is desirable for search engines to fulfil six criteria for their use as a webometric source, they must:

- Have a large database covering as much of the web as possible, as evenly as possible;
- Be up-to-date;
- Have the ability to search for all pages in a particular web space;
- Have the ability to search for all pages that contain links to a particular web space;
- Have the ability to combine search results with Boolean logic;
- Have the ability to provide consistent numeric results.

Whilst these characteristics are desirable, the importance ascribed to each must be seen in relation to the nature of the specific webometric study. This is recognised by Smith (1999b) who mentions that whilst he ignores the effect of outdated links within his own study, such a stance could not be taken if studies were looking at changes in link patterns over time. For certain investigations it may desirable that a search engine does not cover the web 'as evenly as possible', but rather covers those sections of the web of interest to the research question more thoroughly. Whilst it is not possible for a search engine to be totally up-to-date, a webometrician may prefer to have fast changing sections of the web crawled more often at the expense of less volatile areas of the web.

Bar-Ilan's (2005a) more recent discussion of an 'ideal' search engine, reflects growing understanding of the problems inherent within search engines, as well as growing expectations of what a search engine may be able to offer. In addition to Smith's (1999b) criteria Bar-Ilan (2005a) adds:

- Transparency, disclosure, clear documentation.
- Indexing the whole document.

- Response time, accessibility.
- Objectivity – no commercial influences and no influences on the environment.
- All reported results are retrievable.
- Rank, different sorting option.
- Flexible output display.
- Cached results.
- High quality retrieval in languages other than English.
- Accessible application programming interface.
- Wide variety of search modifiers.
- Stemming, truncation, wildcards, case sensitivity, spell check, site collapse.
- Relevance feedback, similar/related pages and searches.
- The ability to combine all the features in a single query.
- Non-textual retrieval capabilities.

As a utopian search engine ideal it seems unlikely that one of the current major search engines will fulfil all of the criteria in the near future. Whilst there may be improvements with regards to some of the requests, others, such as the disclosure of ranking and crawling policies seem unlikely as the search engines battle with the spammers who try to get top rankings for certain keywords by whatever means possible. Whilst the fight against spam has always been claimed as the reason why search engines will not disclose their crawling and ranking algorithms, it has recently been suggested that opening the algorithms to all may be a way of dealing with spam as it would even up the playing field (Lombardi, 2007), it seems more likely however that as personalisation plays more of a part in the ranking of search results the traditional spamming of search engines to be on the front page becomes defunct. Alternatively more open search engines are also under development, but these are currently at the very early stages (e.g., Wikia Search, 2007), and objectivity seems unlikely to be a search engine attribute anytime soon due to the huge influence of a small number of search engines resulting in some web sites buying links to increase their search engine rankings.

Although search engines are not a perfect source for information there have been steps in recent years to make the data within them more accessible through the introduction of Application Programming Interfaces (APIs) as well as new operators. This has enabled them to be used in far more investigations than previously, and more importantly, in a legitimate manner. APIs provide a way for a relatively simple program to automatically retrieve data from a search engine, or other information source, in a manner that may be regulated by the information source. Whilst link analysis investigations have often utilised search engine data in the past they have relied upon search engine scrapers (Larson, 1996; Heimeriks et al., 2003; Heimeriks & van den Besselaar, 2006), programs that download the data from the search engine web site. In reality they are ‘scraping’ the data from the web sites, a practice that is frowned upon as it is not using the search engine for the purposes it is provided, and excessive use may cause damage to the system they are scraping (*see section 3.2 Web crawling ethics*).

Search engines are not perfect, but they may be the only suitable method of data collection, depending on the nature of the investigation. Within his investigation Smith

(1999b) chooses "the nearest to a search engine that satisfied the criteria", which is the best any webometric investigation can do without better alternatives being available. For any webometric investigation relying on search engine data it is necessary to add the caveat, as Vaughan (2004a) has done, that the results are only "robust to the search engine used" to which should also be added: at the time of use.

2.5.3 Data cleaning

Web links have been described as an endorsement (Thelwall, 2004a) and a representation of trust (Davenport & Cronin, 2000). Whilst there are in actuality a far wider range of reasons for link placement (see section 2.5.4.2), it is the belief that the majority represent an endorsement that forms the basis of many link analysis indicators. According to Thelwall (2004a) for the endorsement to be meaningful, and to gain the best results from information science link research, all the links counted should have been created with four characteristics: individually and independently; by humans; with equivalent judgements about the quality of the information in the target page; and links to a site should target pages created by the site owner or somebody else closely associated with the site.

In reality many web links are replicated automatically. Whilst the repetition of navigation bars within a web site was recognised early on in webometric research (Thelwall, 2004a), it is not only self-links that are replicated automatically; acknowledgements may also be replicated automatically. There are also cases of non-replicated automatically generated links; instances have been found of two databases being interlinked with every page on one database having a link to a page on another (Thelwall, 2002d). The problem of replicated outlinks for link analysis has risen with the growth of blogs, which generally copy the site owner's blog roll, the list of other blogs they regularly follow, onto every web page.

As well as the legitimate repetition of web links there are also many spam web sites and blogs generated automatically. Potentially popular URLs, e.g., misspellings of popular URLs, may be created for the purpose of directly driving traffic to a web site that is selling something. Alternatively, web sites may be created to indirectly drive traffic to a site through an increase in inlinks increasing the web site's search engine rankings. The introduction of simple advertising distribution and reward mechanisms such as Google AdSense has also led to an increasing number of web sites which are placed purely for the purpose of attempting to raise money through the clicking of online advertisements.

A partial solution to the problem of self-links and automatically replicated links is the use of Thelwall's (2002d) Alternative Document Models (ADMs). Previous to the introduction of ADMs link analysis investigations counted the links between web pages. Thelwall pointed out that this was not the only way to count links and not necessarily the most appropriate. Rather than counting the number of pages linking to a web site it is possible to count the number of directories, domains, or sub-domains. Utilisation of Thelwall's (2002d) ADMs produces a partial solution to the problem, as it restricts the effect any single web document can have on the number of links a web site receives. However, it is also necessary to investigate the appearance of anomalies manually as web sites build networks of web sites with different domains in an attempt to boost rankings and these will not all be identified through the use of an ADM. Whilst it has been suggested that the ever increasing scrutiny to which citations are subjected reduces specious citations (Davenport & Cronin, 2000), this notion does not seem to be transferable to the web where the worth of links has been recognised, but the sheer size results in a lack of scrutiny and a rise in specious links.

The cleaning of the link data is also dependent on the data source used, for example search engines typically restrict the number of results from a search that may be retrieved. As such cleaning the results from a search engine is fairly limited. Some authors have suggested that search queries may be split through the adding of more restrictive terms when more results are found than may be downloaded (Bar-Ilan & Peritz, 2007; Thelwall, forthcoming). For example, *linkdomain:xxx.co.uk* may produce 1,600 links, i.e., more than a search engine would allow you to retrieve, however, this query may be split into *linkdomain:xxx.co.uk yyyy*, and *linkdomain:xxx.co.uk –yyyy*. The additional term may be a context relevant word extracted from the documents under investigation, or specify a particular section of the web, and additional terms can be added if the number of hits continues to be more than the number of results that are allowed to be viewed. For the maximum number of results Thelwall recommends using a combination of term extraction and TLD splitting. However, such behaviour may be considered unethical by the search engines as it is trying to extract more information from the search engine than the service is designed to provide. Search engines are very protective of their data and if they feel services are being abused they have few qualms about withdrawing the services (Live Search, 2007).

2.5.4 Validation of link analysis

Links may be placed for a wide variety of reasons, and it is therefore necessary to determine whether there is a connection between the links and what is being inferred from them. Thelwall (2004a) suggests that such validation may be partially achieved through correlation tests and partially achieved through classification exercises. Finding a correlation between link data and another quantitative data source relevant to the research question provides support for the validity of the link data to answer the research question. Such correlation tests can only support the validity of the source data, rather than proving it, as a statistically significant correlation doesn't prove causation but may be caused by an additional factor that influences both. Thelwall (2004a) suggests that the correlation exercise is carried out in conjunction with a classification of a random sample of the web links. Causation is provided additional support if the reasons for link placement reflect the research question. For example, if web links are being placed to highlight quality research then it adds support to the notion that web links can be used as an indicator of an institution's research quality. Whilst both correlation tests and classification exercises may add support to the findings of an investigation, any indicators can only ever be non-robust (Thelwall, 2004c).

Whilst validation would seem an important part of an investigation which is going to draw conclusions about the nature of the relationship between different organisations based on the web links between the web sites, or use web links as an evaluative measure, there have been a number of link analysis investigations that have not included any validation. These include exploratory studies into the topological structure of the web, where such validation may be thought unnecessary, as well as studies that have overlooked the subject of validation or have used indirect validation by citing similar validated research.

There have been a number of studies that have looked at how different web sites and TLDs interlink, and how such interlinking may be visualised. Many of these investigations have been carried out without further validation: the suitability of network diagrams to visualise the strength of interconnection between areas of the web (Thelwall, 2001c); manifestations of community structures at various ADMs (Thelwall, 2003b); interlinking between country-level academic web domains (Thelwall & Smith, 2002); the interlinking

between ccTLDs, according to language (Thelwall, Tang, et al., 2003); and the interlinking between academic institutions (Thelwall, 2002b; 2002d). The primary motivation of these studies is to increase the understanding of the patterns in the link structure that later evaluative techniques may build upon, rather than them being evaluative investigations themselves, although such studies could have been enhanced by the incorporation of correlation and classification exercises. For example, investigations into the interlinking between countries would have been enhanced by an investigation into the reasons for link placement as it is possible that the different communities are placing links for different purposes, there are however obvious difficulties in the classification of web links on web pages in numerous different languages.

Early evaluative link analysis investigations also occasionally failed to incorporate any form of validation. In his seminal paper on the calculation of Web Impact Factors (WIF) Ingwersen (1998) proposed that search engine data could be utilised to calculate a web site's impact factor: the sum of inlinks and self-links to a web document, divided by the number of links within the web document. Whilst the results for different countries and domains were discussed, the logical step of comparing with an external quantitative source was not taken, although the WIF has since gone on to form the basis of many later link analysis investigations (e.g., Smith, 1999b; Thomas & Willet, 2000). Most of the evaluative or relational, as opposed to exploratory, link analysis investigations do include at least one of the forms of validation, although it is a relatively small proportion that uses both correlation and classification (e.g., Thelwall, 2001c; Thelwall & Harries, 2004a; Li, 2005).

2.5.4.1 Partially validating link count results through correlation tests

Correlation tests have played a significant role both within the development and application of the WIF and other evaluative measures. It's a natural extension of the concept of the inlink as an endorsement of a web document to investigate a web document's inlinks as an indicator of quality in much the same way as citations have been found to reflect the quality of journals and academic departments (Moed, 2005). As such there have been many investigations into the correlation between a web document's inlinks and a recognised indicator of quality. Primarily such investigations have focused on the higher education institutions within the UK as they offer the opportunity for correlation tests to be carried out with the Research Assessment Exercise (RAE), a qualitative peer review of an institution's submitted research output. Such correlations follow a precedent in citation analysis where citations have previously been found to correlate with RAE findings (e.g., Oppenheim, 1997; Norris & Oppenheim, 2003). Investigations have also looked at correlations between an organisation's inlinks and their financial success in the commercial sector (e.g., Vaughan, 2004a; Vaughan & Wu, 2004), which have a precedent in patent analysis where correlations have been found between increases in company profits and sales and patent citations (Narin, Noma, & Perry 1987).

The Web Impact Factor as proposed by Ingwersen (1998) has been described as the original link metric (Thelwall, 2004a). Drawing on similarities between the web as a network of hyperlinked documents and journals as a network of cited and referenced articles Ingwersen proposed the WIF as an online supplement to traditional impact factors:

> *The dynamic real-time nature of the WWW suggests Web-IFs as a useful supplement to the traditional impact factors when monitoring the status of web locations. They can*

> *be seen as evidence or indicators of the relative attractiveness of countries or research sites on the WWW at a given point in time.* (Ingwersen, 1998, p. 236).

The traditional Journal Impact Factor (JIF) was originally used for monitoring the coverage of the Science Citation Index by identifying the most important journals within a field, and the JIF for a journal in a particular year is defined as the number of citations received by the journal in the year, by all the documents published in the journal in the previous two years, divided by the number of citable documents published in the previous two years (Moed, 2005). By comparison (Ingwersen, 1998):

$$\textit{A web site's WIF} = \frac{\textit{external + self-link pages pointing to the web site}}{\textit{The number of pages in the site}}$$

The principal of the WIF is the substitution of articles as the irreducible units that comprise the journal, with web pages as the irreducible units that comprise a web page. The most striking difference between the two impact factors is that, unlike the JIF, the WIF doesn't only count links from specific years. It is necessary to include time as a variable in the calculation of JIFs as older volumes may not reflect a journal's current status. However, Ingwersen (1998) argues that, unlike citations, which once placed are permanent and therefore the number of citations can only increase, inlinks are transient and may increase, decrease or disappear altogether. The JIF calculates the citations per article as opposed to all the citations to a journal, so that it can act as an indicator of the quality of the journal rather than just reflecting differences in the quantity of documents published by the journal.

There have however been alternative numerators and denominators suggested for the calculation of the WIF. For example, Thomas and Willet (2000) identified six different ways the numerator could be calculated for an academic department:

- All links to the departmental web site.
- Departmental self-links.
- Departmental inlinks.
- The sum of departmental inlinks and departmental self-links.
- Links from within the host institution, but outside the department.
- Links from outside the host university to the department.

Whilst theoretically the sum of the departmental inlinks and departmental self-links should equal all the links to a departmental web site, due to the idiosyncrasies of the AltaVista search engine this wasn't found to be the case and in his original study Ingwersen (1998) recommended the combined totals as the more reliable method. Thomas and Willet's list is by no means exhaustive: inlinks may be calculated from within a particular section of the web, for example, from other academic web sites (e.g., Chen et al., 1998). It not necessary to use the web page as the unit of analysis, the numerator could consist of the number of directories, domains, or web sites that have web links pointing to a particular web site (Thelwall, 2002d). The decision about which links to count will reflect the research question that the link analysis is trying to answer, although as most self-links are placed for navigational reasons rather than as endorsements these are generally not counted (Bar-Ilan, 2004b).

There has also been a departure from the use of the number of web pages as the denominator in the calculation of the WIF as the number of pages is often a reflection of choices in site design rather than the content of the site. The use of full-time equivalent members of staff as the denominator has been found to provide greater correlation with other indicators of organisational quality, such as the RAE (e.g., Li et al., 2003).

There have been a number of studies that have utilised a variation of the WIF with the number of web pages as the denominator and have failed to find a statistically significant correlation with another indicator of quality. Thomas & Willet (2000) failed to find a statistically significant correlation between various versions of the WIF and the RAE rating of Library and Information Science departments in the UK without the removal of Loughborough University as an anomaly with excessive low web links relative to its research ranking. Smith (1999b) didn't find a statistically significant correlation between the external WIF and the number of publications per academic member of staff. Thelwall (2001e) didn't find a statistically significant correlation between the external WIF and an institution's ranking given in the Times Higher Education Supplement. Qiu et al. (2003) failed to find a correlation between Chinese universities' WIFs and the Chinese university rankings. Although some studies have; a correlation was found between a university's RAE using the number of web pages as denominator and the number of inlinks from the academic web in the UK, using data from AltaVista (Thelwall, 2001d). However, when investigating the correlation between the inlinks from different TLDs to university web sites (Thelwall, 2002a) and computer science departments (Li et al., 2003) with the university's/computer science department's research productivity, many of the TLDs that were found to be statistically significant when using the number of FTE members of staff as a denominator were not found to have statistically significant correlations when using the number of pages as a denominator.

It is possible that the lack of correlations in the early investigations (i.e., Smith, 1999b; Thomas & Willet, 2000) was due to the web not being fully developed at that point, and that the lack of correlation within Chinese universities (i.e., Qiu et al., 2003) was due to coverage issues of the Chinese web by the AltaVista search engine. It also seems reasonable, however, to place at least some of the responsibility at the use of the number of pages within a site to normalise the different sized academic web sites as it is the quality that the impact factor is trying to measure rather than the quantity.

The article is the natural unit of investigation within journals as each is distinct from one another, and when a reference is made it is generally made to a specific, identifiable article rather than to a journal generally. The problem with using the number of web pages as the denominator in calculating the WIF is that whilst a link may be placed to point to a single unit, the number of web pages such a unit consists of is dependent on the site structure that is chosen by the individual or the organisation whose site it is placed on. For example, an academic may place an article on the web as a single document with one identifiable URL or alternatively they may choose to place the article over five URLs with a different web page for different sections of the article. Such decisions have a marked effect on the WIF of a web document, in the first scenario if the web document has one inlink the external WIF would be 1, whilst in the second scenario the external WIF would be 0.2, even though both contain the same number of inlinks and the same information. The influence of such decisions means that the WIF may end up measuring a web site's tendency to produce large pages rather than the impact of its information.

Some studies have not included the number of pages in a site as a denominator, but rather have looked at the number inlinks in comparison to recognised indicators of quality.

Chen et al.'s (1998) investigation into the inlinks to Scottish universities' computer science departments from other Scottish universities' computer science departments found a correlation with the departments' research funding levels, teaching quality assessment, and RAE, whilst a high student-staff ratio and a high proportion of teaching funding tended to have fewer inlinks. Whilst Qiu et al. (2003) failed to find a correlation between an institution's Chinese University Rating and its WIF, they did find a correlation between the Chinese University Rating and the number of inlinks. The statistically significant correlations would suggest that the computer science departments are sufficiently similar for any effect size difference may have on a rating to be negligible. It would be expected that the differences in the size and scope of academic institutions would be reflected in the number of inlinks the Chinese universities received, irrespective of the quality of the university. For example it would be expected that a small highly specialised university of international merit would receive fewer links than a large multidisciplinary university, merely because the people it relates to are going to be fewer; within the UK we would not expect to find the School of Oriental and African Studies to have more inlinks than the University of Wolverhampton in spite of its far higher research ranking. That the inlinks correlate with the universities' ranking raises questions about what is being measured in the ranking system, the institutions included in the ranking system, and possibly the uneven coverage of the Chinese web space by the AltaVista search engine. Whilst Vaughan (2004a) found a statistically significant correlation between an organisation's number of inlinks and revenue for technology companies in the US, it was recognised that this was likely to be partially attributable to the size of the institutions rather than necessarily reflecting the efficiency of the organisations, and as such the correlation between the number of inlinks was investigated with the organisational revenue per employee.

It has been suggested that using the number of full-time equivalent (FTE) members of staff as a denominator is a better way of normalising for the size of an organisation rather than the number of web pages, and the statistically significant correlations that have been found seem to back this up. Using the FTE employee as a denominator statistically significant correlations have been found between a range of inlinks and indicators of academic success: inlinks from research related pages on other universities' research related web pages and a university's RAE rating, as well as inlinks from the academic web in the UK (Thelwall, 2001d); inlinks to university web sites from various different TLDs and the RAE ratings (Thelwall, 2002a); inlinks to UK universities from other UK universities and a university's RAE when counting the inlinks from web pages, directories, domains, or universities, as well as when counting links between pairs of universities (Thelwall, 2002d); inlinks to UK universities from UK universities to be correlated with an institutions RAE rating (Thelwall, 2002e); inlinks from different TLDs and a computer science department's RAE rating (Li et al., 2003); and domain inlinks to domain documents within a university and the university's RAE (Thelwall & Harries, 2004b). The number of FTE members of staff is thought to be a better denominator as the result can not be attributed solely to a decision on web site design. It should be noted that whilst there are a number of statistically significant results, there have been studies where correlations have not been found: Qiu et al.'s (2003) investigation failed to find a correlation between WIFs that normalise by the number of staff.

It is unsurprising, as the WIF was originally derived from the JIF, to also find the relationship between the WIF and the JIF to be the focus of a number of investigations. Whilst some investigations have investigated the WIF of electronic journals (Smith , 1999; Vaughan & Thelwall, 2003b), others have compared the WIF and the Publication Impact Factor for

academic departments (the number of publications for a department multiplied by the ISI citations impact for the particular field) (Tang & Thelwall, 2003a).Whereas when Smith (1999) investigated electronic journals he found them poorly covered within the ISI citation databases, by the time of Vaughan and Thelwall's (2003b) investigation there were more electronic journals available on the web, allowing for correlations between WIFs and JIFs to be investigated. They found that there were statistically significant correlations.

Tang and Thelwall's (2003a) investigation into the correlations between the Publication Impact Value of departments and a department's international inlinks found a statistically significant correlation for both chemistry and psychology (there were not enough links to history departments to permit statistical investigations). Whilst there was also found to be a correlation between international inlinks per faculty member and total publication impact per faculty member for the Psychology departments, this was not found for the Chemistry departments.

The popularity of investigating inlinks may be attributed to the fact that they are generally considered to be an emergent property of numerous other people's opinions, and as such are more difficult to manipulate than indicators such as an organisations outlinks. Nonetheless outlinks have also been the focus of link analysis, as has a web site's connectivity.

Thelwall (2003c) introduced two new metrics for university web sites, the web use factor (WUF) and the connectivity factor (WCF). The WUF is defined as the number of outlinks from a university to an area of interest divided by the number of FTE academic staff, and is proposed as a logical companion to the WIF. As both the WIF and the WUF are susceptible to a small number of pages having huge numbers of outlinks or inlinks, as is the JIF (Moed, 2005), Thelwall proposes the WCF for a university as a more stable metric: the minimum of, the number of links originating from a page of one university targeting a page of another university and the number of links at the other university targeting a page at the first university, summed over all the universities in the area, divided by the FTE members of staff. Within the investigation Thelwall found the WIF, the WCF, and the WUF to be correlated with an organisation's RAE rating.

Whilst finding correlations between the number of links to an organisation's web site and an external indicator of organisational quality may be aesthetically pleasing it does not in itself mean there is a direct link between the two; web links are placed for a whole host of reasons, and few of them are likely to be explicitly identifying high quality research. There are a variety of different reasons why some web sites are more likely to have inlinks than others, and care needs to be taken so that web sites are being compared on an equal footing. The necessity of comparisons being carried out between homogenous organisations however causes difficulties in the use of web links as a quantitative indicator of organisations within knowledge-based innovation systems, which are by definition heterogeneous. Investigating the roles of organisations from different sectors of society, is unlikely to be reducible to 'organisation X is more triple-helixed than organisation Y' based on inlinks alone. In much the same way that more inlinks need to be normalised according to the size of an organisation, they also need to be normalised for all the other factors that contribute to the links an organisation may receive, e.g., the technological focus of the organisation.

2.5.4.2 Partially validating the interpretation of the results through a link classification exercise

For link indicators to have any credibility it is necessary for there to be a level of understanding as to how and why they are placed, a 'theory of linking'. As with so much of link analysis there is a precedent within citation analysis where there has long been the search for a 'theory of citation'. The development of the citation indexes in the 1960s enabled statistical analysis of the citation process, and it was quickly recognised that there were discernable patterns, which as Kaplan (1965) stated: "suggests some underlying, but as yet ill-defined, set of norms governing the behaviour of scientists communicating their results in the periodical literature". Such discernable patterns have also been recognised on the web (Egghe, 2000; Adamic & Huberman, 2001).

Moed (2005) recognises many different approaches towards a theory of citation: the physical; the sociological; the psychological; the historical; and the information and communication science. Whilst some of these approaches recognise the place of citation classification as a step towards a theory of citation, others have little room for such exercises:

> *It is as if a physicist would strive for creating a framework of thermodynamics by making a 'theory' on the behaviour of individual molecules* (van Raan, 1998, in Moed, 2005, p.207).

The physical approach to citation, of which van Raan is an advocate, rejects classification of individual citations in favour of statistical approaches, where the properties of collections of citations are investigated. The sociological, psychological and historical however all have a place for classification exercises as attempts are made to understand the relationships between not only the documents, but also the document authors, at an individual level. The sociological approach looks at the role of citations within the framework of expected norms of behaviour of the group, the psychological approach looks at references from the perspective of why people reference the way they do, whilst the historical approach looks at what citations can tell us about the flows of ideas in the development of a science or technology. Moed (2005) notes that the information science approach towards a theory of citation borrows from both the physical and sociological approaches focusing both on information as an entity whose flows may be modelled and as the manifestation of social structures.

The appropriateness of a classification of citations is based primarily on the type of investigation. It is not enough to accept broad conclusions based on macro statistical investigations when looking at a micro level and equally micro level investigations should not be used to dismiss certain macro level investigations. For example, although the structure of the web is used as part of Google's ranking algorithm for the promotion of high quality web sites (Brin & Page, 1998) with great success, we can not infer that the links are placed to indicate high quality sites, and equally finding that a proportion of those links are placed for negative reasons does not detract from the quality of the sites that are found utilising the link structure. Within Thelwall's (2004a) link analysis methodology he states that finding a proportion of insignificant links is not necessarily a problem as long as the proportion is not too high, a variable he recognises as specific to each investigation.

Citation classification may be carried out in two ways: either through a content analysis of the citing article (e.g., Moravcsik & Murugesan, 1975) or through questioning a paper's author (e.g., Brooks, 1985; 1986; Leydesdorff & Amsterdamska, 1990; Case & Higgins, 2000; White, Wellman & Nazar, 2004). Whilst the classification of an article's citations by content analysis is limited to the textual reasons for a citation being placed, these

do not necessarily correspond with the subjective reasons (Leydesdorff & Amsterdamska, 1990). Also such investigation must necessarily ignore the reasons citations are not included (MacRoberts & MacRoberts, 1988). The sending of questionnaires to authors is likely to increase our understanding of the decisions made in placement of citations, however, the inclusion of citations in an article is likely to reflect a myriad of choices that an author may not be consciously aware of, or may not be willing to be open about. It is also not only the author whose input needs to be taken into consideration, whilst the role of other parties in the placement of patent citations is recognised (Meyer, 2000b), it is also necessary to recognise the role of referees and editors in the citation process in science articles. A referee may reject a work on account of what is perceived to be an insufficient literature review, whilst editors may encourage journal self-citations.

There has been a number of citation classification schemes proposed for the analysis of citations (Cronin, 1984), reflecting not only the broad range of reasons why citations are placed, e.g., giving credit, highlighting previous work, providing authority, as well as social factors (Campanario, 2003), but also reflecting the contribution of a cited document to a paper (e.g., Chubin & Moitra, 1975). Although much of citation theory has focused on the reward notion of citations, Cozzens (1989) argues that citations are primarily placed for rhetorical reasons and that the reward is secondary.

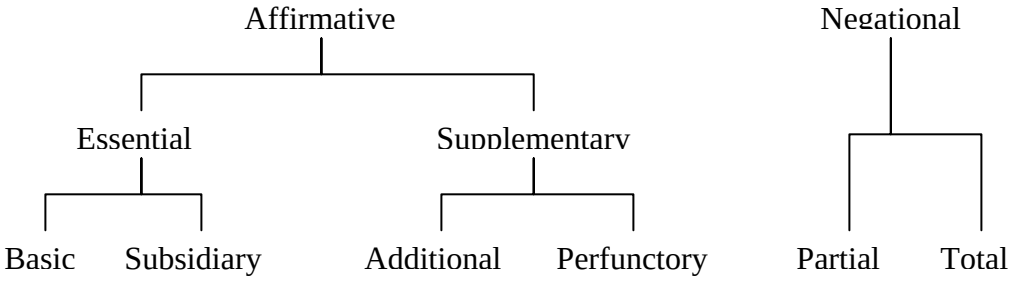

Figure 2-3 Chubin & Moitra's (1975) six classification classes.

Whilst there are similarities between citation and link networks it is not necessarily appropriate to use a citation typology for the classification of web links. Chubin and Moitra's (1975) proposed classification typology (see Figure 2-3) is a mutually exclusive adaptation of Moravcsik and Murugesan's earlier typology which allowed citations to be classified into more than one category; it provides a framework for citations as an indicator of a paper's worth to be investigated. The creation of a similar classification scheme for determining the contribution of a link to a web page would be far more difficult to assess; few links are clearly positive or negative (Smith, 2003) and web pages are placed for a far wider range of reasons, which are often ill-defined and personal. It would be impossible to classify the links on someone's homepage as supplementary or essential when it is used purely as a list of personal bookmarks without interviewing the web page owner as to the purpose of the web page, and their use of the target web page. Unlike journal articles and the embedded citations, web pages and their links are not necessarily attributable to any particular person and as such the classifications have focused on the information discernable from content analysis of the web pages rather than through questionnaires.

Most of the link classifications have also been constructed for the answering of a particular research question due to the broad range of reasons people use the web, as well as the wide range of web-based research questions, which makes a widely agreed typology for links unlikely (Harries et al., 2004) in much the same way as Garfield (1998) has previously commented that there will probably never be a complete typology of citation behaviour. Haas and Grams (1998) suggested a broad link-classification scheme consisting of three categories: navigation around a web site; expanding on the details in the anchor text; pointing to further resources; or for miscellaneous reasons that included courtesy links and advertisements. The limited nature of the classification is likely to be a reflection of the initial web link selection. The study used AltaVista's surprise link feature, which is no longer available. It seems likely that the majority of pages returned were commercial web sites which have a tendency to have a limited number of outlinks (Shaw, 2001), and this is reflected in the detailed description in the types of navigational links that may occur whereas there is little further breakdown of the different types of resources that may be linked to. It is unsurprising therefore to find that with most of the link analysis investigations focusing on the academic community, that they have created their own typologies to reflect the link placement on academic web sites.

The navigational link was also recognised in Thelwall's (2003d) investigation into the reason link placement as one of four roles of web links that have no precedent in citations. Thelwall classified 100 web links that were found on one university web page and pointed to a different university's homepage and identified navigational, ownership, social, and gratuitous links as having motivations that are "different from those normally associated with citations". Whilst the motivations are different from those normally associated with citations, they do nonetheless have bibliometric precedents as Thelwall acknowledges. General navigational links have a precedent in citations to set the background of a study; ownership links have a precedent in the co-authorship and acknowledgements; social links have been recognised as a factor in citation motivation. Whilst Thelwall does not mention a bibliometric precedent for the gratuitous link, a link with little or no value to the viewer of the page, it seems likely that there are many examples within the traditional journal where authors may shoehorn in previous articles they may have written on the flimsiest of bases to increase their personal number of citations.

The similarities between citations and web links have been the starting point of a number of the classification exercises. Kim (2000) compared the motivations for hyperlinking within electronic articles and the traditional citer motivations. Thelwall, Harries, et al. (2003), as well as Harries, Wilkinson et al. (2004), investigated academic subject interlinking in an attempt to establish the suitability of using web links for science mapping, a traditional application of citation analysis. Nevertheless links and citations are different, and although there are without doubt many lessons to learn from citation analysis, it is necessary that care is taken not to take the analogy too far.

As well as links being classified for comparison with citations, links have also been used for establishing substantive-WIFs, impact factors that count the relevant links (Smith, 2003), and to determine the type of information that may be extracted from them: Thelwall (2001b) classified the links on commercial web sites; Vasileiadou & van den Besselaar (2006) investigated the reasons for link placement on the web sites of scientific intermediaries in the Netherlands; and Bar-Ilan (2004a; 2005b) investigated the reasons for link placement between academic institutions. Despite investigating different sections of the web each of these classifications has found a proportion of the web links that reflect some form of collaboration between the interlinked institutions.

It is worth discussing the investigation of Vasileiadou and van den Besselaar (2006) in more detail as they focus specifically on the possibility of using web links to investigate the collaborations between organisations, and it is the relationships between organisations that are of interest in an investigation into knowledge-based innovation systems. Their study analysed outlinks of eight scientific intermediaries, collected through a bespoke web crawler that crawled the web sites four levels deep. They classified the outlinks according to whether there was a practical collaboration between the originating organisation and the target organisations, which was determined through analysis of relevant documents found online as well as the information found within the source page of the outlink. The outlinks were classified according to whether they reflected a collaborative relationship or not, with the collaboration category being further sub-divided in three non-exclusive categories: collaboration involving a financial transaction; collaboration in the form of one company representing another; and collaboration in the form of the two organisations working together. The study found that overall 33% of the coded links were to collaborating organisations, with the majority of the links on the front pages of a web site pointing to a collaborator (61.4%). It is difficult to assess the practicalities of the classification scheme as only one classifier was used, which whilst suitable for an exploratory study gives no indication of whether another classifier would apply it in the same way.

Whilst Vasileiadou and van den Besselaar's (2006) work supports the idea that web links may be of use for investigating collaborative relationships between organisations, it is important to take into consideration that it only looked at a very small selection of one particular type of organisation which has the purpose of building bridges between the science of academia and organisations from other sectors. There are many other organisations with different purposes with different expectations of their web sites, and the findings do not necessarily transfer to different types of organisation.

2.6 Summary

The vast size of the web and the collocation of a lot of previously disparate organisations provides an important source of information about the relationships between different organisations that the literature shows has not yet been fully exploited. Up until now webometric investigations have either been macro-level with little interest about what is represented by the individual links, or micro-level investigations that have focused primarily on the academic sector. The micro-level is an important area of investigation when looking at the manifestations of knowledge-based innovation systems on the web as it is necessary to understand the differences between the linking behaviour of different organisations in different sectors.

Although the link analysis methodology, as described by Thelwall (2004a), emphasises the importance of both correlation and classification for the validation of conclusions based on link counts, the diverse range of organisations involved within knowledge-based innovation systems makes validation through correlation impossible as it is only suitable for homogenous organisations. Classification of web links has shown that although the majority that are placed do not reflect a relationship between the linked organisations, studies have found that there is a significant proportion that do. It is important to understand whether this information is different to that contained within traditional bibliometric sources such as science papers and patents.

3 Preliminary investigations

3.1 Introduction

This chapter reports on five, previously published, preliminary studies that establish the suitability of the web for investigations into knowledge-based innovation systems. The preliminary investigations play a crucial role within webometric investigations, they determine the feasibility of the final investigation and validity of such an investigation.

The first study, *Web crawling ethics revisited: Cost, privacy and denial of service* (Thelwall & Stuart, 2006), looks at the ethical implications of the data collection process. This is an essential initial step within any investigation, but one that has been missing from previous webometric studies which have focused primarily on the ability of the different data collection methodologies to provide the data needed.

This is followed by two studies that investigate the reasons links are placed on academic web sites to non-academic web sites: *What can university-to-government web links reveal about university-government collaboration?* (Stuart & Thelwall, 2005); *University outlinks: What do links to different domains represent?* (Stuart et al., 2007). With previous investigations focusing primarily on the reasons for link placement between different academic institutions it was necessary to determine whether, as with previous studies, a significant proportion of web links were found to reflect collaboration between organisations.

Finally, the last two preliminary investigations are small scale pilot studies looking at the interlinking between institutions from the university, industry, and government sectors: *Investigating Triple Helix relationships using URL citations: A case study of the UK West Midlands automobile industry* (Stuart & Thelwall, 2006); *University-industry-government relationships manifested through MSN reciprocal links* (Stuart & Thelwall, 2007). The initial pilot study tests the feasibility of using the recommended ethical data collection techniques and what is shown in a small-scale investigation. The second of the two pilot studies reflects changes in the tools available to investigate linking practices and reappraises the potential of a large scale investigation.

3.2 Web crawling ethics revisited: Cost, privacy and denial of service

3.2.1 Introduction

Web crawlers, programs that automatically find and download web pages, have become essential to the fabric of modern society. This strong claim is the result of a chain of reasons: the importance of the web for publishing and finding information; the necessity of using search engines like Google to find information on the web; and the reliance of search engines on web crawlers for the majority of their raw data, as shown in Figure 3-1(Brin & Page, 1998; Chakrabarti, 2003). The societal importance of commercial search engines is emphasized by Van Couvering (2004), who argues that they alone, and not the rest of the web, form a genuinely new *mass* media.

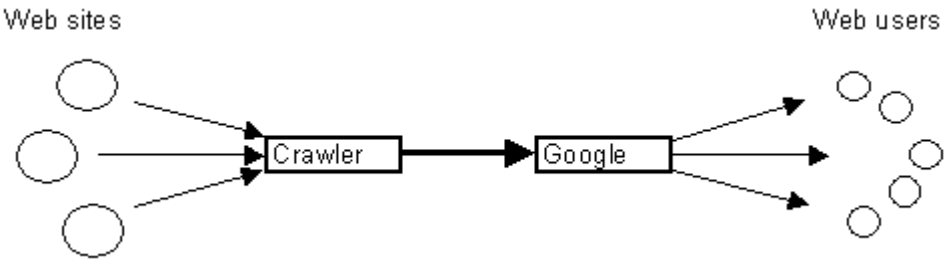

Figure 3-1 Google-centered information flows: the role of web crawlers

Web users do not normally notice crawlers and other programs that automatically download information over the Internet. Yet, in addition to the owners of commercial search engines, they are increasingly used by a widening section of society including casual web users, the creators of email spam lists and others looking for information of commercial value. In addition, many new types of information science research rely upon web crawlers or automatically downloading pages. Web crawlers are potentially very powerful tools, with the ability to cause network problems and incur financial penalties to the owners of the web sites crawled. There is, therefore, a need for ethical guidelines for web crawler use. Moreover, it seems natural to consider together ethics for all types of crawler use, and not just information science research applications such as those referenced above.

The robots.txt protocol (Koster, 1994) is the principal set of rules for how web crawlers should operate. This only gives web site owners a mechanism for stopping crawlers from visiting some or all of the pages in their site. Suggestions have also been published governing crawling speed and ethics (e.g., Koster, 1993; 1996), but these have not been formally or widely adopted, with the partial exception of the 1993 suggestions. Nevertheless, since network speeds and computing power have increased exponentially, Koster's 1993 guidelines need reappraisal in the current context. Moreover, one of the biggest relevant changes between the early years of the web and the present is in the availability of web crawlers. The first crawlers must have been written and used exclusively by computer scientists who would have been aware of network characteristics, and could have easily understood crawling impact. Today, in contrast, free crawlers are available online. In fact there are site downloaders or offline browsers that are specifically designed for general users to crawl individual sites, (there were 31 free or shareware downloaders listed in tucows.com on March 4, 2005, most of which were also crawlers). A key new problem, then, is the lack of network knowledge by crawler owners. This is compounded by the complexity of the Internet, having broken out of its academic roots, and the difficulty to obtain relevant cost information (see below). In this paper, we review new and established moral issues in order to provide a new set of guidelines for web crawler owners. This is preceded by a wider discussion of ethics, including both computer and research ethics, in order to provide theoretical guidance and examples of more established related issues.

3.2.2 Introduction to ethics

The word 'ethical' means, 'relating to, or in accord with, approved moral behaviors' (Chambers, 1991). The word 'approved' places this definition firmly in a social context. Behavior can be said to be ethical relative to a particular social group if that group would approve of it. In practice, although humans tend to operate within their own internal moral code, various types of social sanction can be applied to those employing problematic

behavior. Formal ethical procedures can be set up to ensure that particular types of recurrent activity are systematically governed and assessed, for example, in research using human subjects. Seeking formal ethical approval may then become a legal or professional requirement. In other situations ethical reflection may take place without a formal process, perhaps because the possible outcomes of the activity might not be directly harmful, although problematic in other ways. In such cases it is common to have an agreed written or unwritten ethical framework, sometimes called a code of practice or a set of guidelines for professional conduct. When ethical frameworks or formal procedures fail to protect society from a certain type of behavior, it has the option to enshrine them in law and apply sanctions to offenders.

The founding of ethical *philosophy* in Western civilization is normally attributed to ancient Greece and Socrates (Arrington, 1998). Many philosophical theories, such as utilitarianism and situation ethics, are relativistic: what is ethical for one person may be unethical for another (Vardy & Grosch, 1999). Others, such as deontological ethics, are based upon absolute right and wrong. Utilitarianism is a system of making ethical decisions, the essence of which is that "an act is right if and only if it brings about at least as much net happiness as any other action the agent could have performed; otherwise it is wrong." (Shaw, 1999, p.10). Different ethical systems can reach opposite conclusions about what is acceptable: from a utilitarian point of view car driving may be considered ethical despite the deaths that car crashes cause but from a deontological point of view it could be considered unethical. The study of ethics and ethical issues is a branch of philosophy that provides guidance rather than easy answers.

3.2.3 Computer ethics

The philosophical field of computer ethics deals primarily with professional issues. One important approach in this field is to use social contract theory to argue that the behavior of computer professionals is self-regulated by their representative organizations, which effectively form a contract with society to use this control for the social good (Johnson, 2004), although the actual debate over moral values seems to take place almost exclusively between the professionals themselves (Davis, 1991). A visible manifestation of self-regulation is the production of a code of conduct, such as that of the Association for Computing Machines (ACM, 1992). The difficulty in giving a highly prescriptive guide for ethical computing can be seen in the following very general important advice, "One way to avoid unintentional harm is to carefully consider potential impacts on all those affected by decisions made during design and implementation" (ACM, 1992).

There seems to be broad agreement that computing technology has spawned genuinely new moral problems that lack clear solutions using exiting frameworks, and require considerable intellectual effort to unravel (Johnson, 2004). Problematic areas include: content control including libel and pornography (Buell, 2000); copyright (Borrull & Oppenheim, 2004); deep linking (Fausett, 2002); privacy and data protection (Carey, 2004; Reiman, 1995; Schneier, 2004); piracy (Calluzzo & Cante, 2004); new social relationships (Rooksby, 2002); and search engine ranking (Introna & Nissenbaum, 2000; Vaughan & Thelwall, 2004; Search Engine Optimization Ethics, 2002). Of these, piracy is a particularly interesting phenomenon because it can appear to be a victimless crime and one that communities of respectable citizens would not consider to be unethical, even though it is illegal (Gopal, Sanders, Bhattacharjee, Agrawal, & Wagner, 2004). Moreover new ethics have been created that advocate illegal file sharing in the belief of creating a better, more open society (Manion & Goodrum, 2000).

Technology is never inherently good or bad; its impact depends upon the uses to which it is put as it is assimilated into society (du Gay, Hall, Janes, Mackay, & Negus, 1997). Some technologies, such as medical innovations, may find themselves surrounded at birth by a developed ethical and/or legal framework. Other technologies, like web crawlers, emerge into an unregulated world in which users feel free to experiment and explore their potential, with ethical and/or legal frameworks later evolving to catch up with persistent socially undesirable uses. Two examples below give developed illustrations of the latter case.

The fax machine, which took off in the eighties as a method for document exchange between businesses (Negroponte, 1995), was later used for mass marketing. This practice cost the recipient paper and ink, and was beyond their control. Advertising faxes are now widely viewed as unethical but their use has probably died down not only because of legislation which restricted its use (HMSO, 1999), but because they are counterproductive; as an unethical practice they give the sender a bad reputation.

Email is also used for sending unwanted advertising, known as spam (Wronkiewicz, 1997). Spam may fill a limited inbox, consume the recipient's time, or be offensive (Casey, 2000). Spam is widely considered unethical but has persisted in the hands of criminals and maverick salespeople. Rogue salespeople do not have a reputation to lose nor a need to build a new one and so their main disincentives would presumably be personal morals, campaign failure or legal action. It is the relative ease and ultra-low cost of bulk emailing that allows spam to persist, in contrast to advertising faxes. The persistence of email spam (Stitt, 2004) has forced the hands of legislators in order to protect email as a viable means of communication (www.spamlaws.com). The details of the first successful criminal prosecution for Internet spam show the potential rewards on offer, with the defendant amassing a 24 million dollar fortune (BBCNews, 4/11/2004). The need to resort to legislation may be seen as a failure of both ethical frameworks and technological solutions, although the lack of national boundaries on the Internet is a problem: actions that do not contravene laws in one country may break those of another.

3.2.4 Research ethics

Research ethics are relevant to a discussion of the use of crawlers, to give ideas about what issues may need to be considered, and how guidelines may be implemented. The main considerations for social science ethics tend to be honesty in reporting results and the privacy and well-being of subjects (e.g., Penslar, 1995). In general, it seems to be agreed that researchers should take responsibility for the social consequences of their actions, including the uses to which their research may be put (Holdsworth, 1995). Other methodological-ethical considerations also arise in the way in which the research should be conducted and interpreted, such as the influence of power relationships (Williamson, & Smyth, 2004; Penslar, 1995).

Although many of the ethical issues relating to information technology are of interest to information scientists, it has been argued that the focus has been predominately on professional codes of practice, the teaching of ethics, and professional dilemmas, as opposed to research ethics (Carlin, 2003). The sociology-inspired emerging field of Internet research (Rall, 2004a) has developed guidelines, although they are not all relevant since its research methods are typically qualitative (Rall, 2004b). The fact that there are so many different environments (e.g., web pages, chatrooms, email) and that new ones are constantly emerging means that explicit rules are not possible, instead broad guidelines that help researchers to

appreciate the potential problems are a practical alternative. The Association of Internet Researchers has put forward a broad set of questions to help researchers come to conclusions about the most ethical way to carry out Internet research (Ess & Committee, 2002), following an earlier similar report from the American Association for the Advancement of Science (Frankel & Siang, 1999). The content of the former mainly relates to privacy and disclosure issues and is based upon considerations of the specific research project and any ethical or legal restrictions in place that may already cover the research. Neither allude to automatic data collection.

Although important aspects of research are discipline-based, often including the expertise to devise ethical frameworks, the ultimate responsibility for ethical research often lies with universities or other employers of researchers. This manifests itself in the form of university ethics committees (e.g., Jankowski & van Selm, 2001), although there may also be subject specialist subcommittees. In practice, then, the role of discipline or field-based guidelines is to help researchers behave ethically and to inform the decisions of institutional ethics committees.

3.2.5 Web crawling issues

Having contextualized ethics from general, computing, and research perspectives, web crawling can now be discussed. A web crawler is a computer program that is able to download a web page, extract the hyperlinks from that page and add them to its list of URLs to be crawled (Chakrabarti, 2003). This process is recursive, so a web crawler may start with a web site home page URL and then download all of the site's pages by repeatedly fetching pages and following links. Crawling has been put into practice in many different ways and in different forms. For example, commercial search engines run many crawling software processes simultaneously, with a central coordination function to ensure effective web coverage (Chakrabarti, 2003; Brin & Page, 1998). In contrast to the large-scale commercial crawlers, a personal crawler may be a single crawling process or a small number, perhaps tasked to crawl a single web site rather than the 'whole web'.

As computer programs, many crawler operations are under the control of programmers. For example, a programmer may decide to insert code to ensure that the number of URLs visited per second does not exceed a given threshold. Other aspects of a crawler are outside of the programmer's control. For example, the crawler will be constrained by network bandwidth, affecting the maximum speed at which pages can be downloaded.

The use of web crawlers by a wider segment of the population, rather than being the preserve of computer science researchers, affects the kinds of issues that are relevant. Table 3-1 records some user types and the key issues that particularly apply to them, although all of the issues apply to some extent to all users. Note that social contract theory could be applied to the academic and commercial computing users, but perhaps not to non-computing commercial users and not to individuals. These latter two user types would therefore be more difficult to control through informal means.

Table 3-1 Academic (top) and non-academic uses of crawlers

User/use	Issues
Academic computing research developing crawlers or search engines. (Full-scale search engines now seem to be the exclusive domain of commercial companies, but crawlers can still be developed as test beds for new technologies.)	High use of network resources. No direct benefits to owners of web sites crawled. Indirect social benefits.
Academic research using crawlers to measure or track the web (e.g., webometrics, web dynamics).	Medium use of network resources. Indirect social benefits.
Academic research using crawlers as components of bigger systems (e.g., Davies, 2001).	Variable use of network resources. No direct benefits to owners of web sites crawled. Indirect social benefits.
Social scientists using crawlers to gather data in order to research an aspect of web use or web publishing.	Variable use of network resources. No direct benefits to owners of web sites crawled. Indirect social benefits. Potential privacy issues from aggregated data.
Education, for example the computing topic of web crawlers and the information science topic of webometrics.	Medium use of network resources from many small-scale uses. No direct benefits to owners of web sites crawled. Indirect social benefits.
Commercial search engine companies.	Very high use of network resources. Privacy and social accountability issues.
Competitive intelligence using crawlers to learn from competitors' web sites and web positioning.	No direct benefits to owners of web sites crawled, and possible commercial disadvantages.
Commercial product development using crawlers as components of bigger systems, perhaps as a spin-off from academic research.	Variable use of network resources. No direct benefits to owners of web sites crawled.
Individuals using downloaders to copy favorite sites.	Medium use of network resources from many small-scale uses. No form of social contract or informal mechanism to protect against abuses.
Individuals using downloaders to create spam email lists.	Privacy invasion from subsequent unwanted email messages. No form of social contract or informal mechanism to protect against abuses. Criminal law may not be enforceable internationally.

There are four types of issue that web crawlers may raise for society or individuals: denial of service, cost, privacy and copyright. These are defined and discussed separately below.

3.2.5.1 Denial of service

A particular concern of web masters in the early years of the Internet was that web crawlers may slow down their web server by repeatedly requesting pages, or may use up limited network bandwidth (Koster, 1993). This phenomenon may be described as denial of service, by analogy to the denial of service attacks that are sometimes the effect of computer viruses. The problem is related to the speed at which services are requested, rather than the overall volume. Precise delays can be calculated using queuing theory, but it is clear that any increase in the use of a limited resource from a random source will result in deterioration in service, on average.

A server that is busy responding to robot requests may be slow to respond to other users, undermining its primary purpose. Similarly a network that is occupied with sending

many pages to a robot may be unable to respond quickly to other users' requests. Commercial search engine crawlers, which send multiple simultaneous requests, use checking software to ensure that high demands are not placed upon individual networks.

The design of the web pages in a site can result in unwanted denial of service. This is important because a web site may have a small number of pages but appear to a crawler as having a very large number of pages because of the way the links are embedded in the pages (an unintentional 'spider trap'). If there is a common type of design that causes this problem then the crawler could be reprogrammed to cope with it. Given the size of the web, however, it is practically impossible for any crawler to be able to cope with all of the potentially misleading sites, and so some will have their pages downloaded many times by the crawler, mistakenly assessing them all to be different.

3.2.5.2 Cost

Web crawlers may incur costs for the owners of the web sites crawled by using up their bandwidth allocation. There are many different web hosts, providing different server facilities, and charging in different ways. This is illustrated by costs for a sample of ten companies offering UK web hosting, which were selected at random (Table 3-2). Something that was noticed when searching for information on pricing structures was the difficulty in finding the information. This was especially true for finding out the consequences of exceeding the monthly bandwidth allocation, which was often written into the small print, usually on a different page to the main charging information.

Different hosts were found to allow a wide variety of bandwidths, with none of those examined allowing unlimited bandwidth. Some other hosts do, but this option is usually reserved for those with premium packages. The consequences of exceeding the monthly bandwidth ranged from automatically having to pay the excess cost, to having the web site disabled.

Table 3-2 Bandwidth costs for a random selection of 10 sites from the yahoo.co.uk directory

Web host	Minimum monthly bandwidth	Web space as a % of bandwidth.	Cost of excess use
www.simply.com	1000 Mb	1-infinity	unclear
www.giacomworld.com	2 Gb	4.88%	unclear
www.cheapdomainnames.net	2 Gb	3.91-4.88%	2p/Mb
www.inetc.net	2,000 Mb	1.25%-10%	2p/Mb
www.hubnut.net	1,000 Mb	10%	5p/Mb
www.webfusion.co.uk	7 Gb	8.37%-infinity	5p/Mb
www.kinetic-internet.co.uk	1 Gb	7.81%-10.24%	1p/Mb
www.services-online.co.uk	0.25 Gb	32.55- 39.06%	5p/Mb
www.databasepower.net	500 Mb	2.44-6.51%	£3/Gb (0.29p/Mb)
www.architec.co.uk	1.5 Gb	6.51-20%	£12/Gb (1.17p/Mb)

As some sites claim to offer 'unlimited' web space, alongside a limited bandwidth, it is clear that problems can be caused quite quickly through the crawling of an entire web site. Even for

those hosts that restrict the amount of web space available, downloading all pages could make up a significant percentage of the total bandwidth available.

3.2.5.3 Privacy

For crawlers, the privacy issue appears clear-cut because everything on the web is in the public domain. Web information may still invade privacy if it is used in certain ways, principally when information is aggregated on a large scale over many web pages. For example, a spam list may be generated from email addresses in web pages, and Internet 'White Pages' directories may also be automatically generated. Whilst some researchers advocate the need for informed consent (Lin & Loui, 1998) others disagree and emphasize the extreme complexity of the issue (Jones, 1994).

3.2.5.4 Copyright

Content crawlers ostensibly do something illegal: they make permanent copies of copyright material (web pages) without the owner's permission. Copyright is perhaps the most important legal issue for search engines (Sullivan, 2004). This is a particular problem for the Internet Archive (http://www.archive.org/), which has taken the role of storing and making freely available as many web pages as possible. The Archive's solution, which is presumably the same as for commercial search engines, is a double opt-out policy (Internet Archive, 2005). Owners can keep their site out of the archive using the robots.txt mechanism (see below), and non-owners who believe that their copyright is infringed by others' pages can write to the Archive, stating their case to have the offending pages removed. The policy of opt-out, rather than opt-in, is a practical one but is legally accepted (or at least not successfully challenged in the U.S. courts) because the search engines have not been shut down. Google keeps a public copy (cache) of pages crawled, which causes a problem almost identical to that of the Archive. Google has not been forced to withdraw this service (Olsen, 2003) and its continued use does not attract comment so it appears to be either legal, or not problematic.

3.2.6 The robots.txt protocol

The widely followed 1994 robots.txt protocol was developed in response to two basic issues. The first was protection for site owners against heavy use of their server and network resources by crawlers. The second was to stop web search engines from indexing undesired content, such as test or semi-private areas of a web site. The protocol essentially allows web site owners to place a series of instructions to crawlers in a file called robots.txt to be saved on the web server. These instructions take the form of banning the whole site or specific areas (directories) from being crawled. For example, if the text file contained the information below, then crawlers (user-agents) would be instructed not to visit any page in the test directory or any of its subdirectories.

```
User-agent *
Disallow /test/
```

A 'well behaved' or ethical crawler will read the instructions and obey them. Despite attempts since 1996 to add extra complexity to the robots.txt convention (Koster, 1996) it has remained unchanged. Some of the big search engine owners have extended the robots.txt protocol to include a crawl delay function, however (AskJeeves, 2004; Yahoo!, 2004). Whilst such an

extension is useful, it is important that it is incorporated as a standard protocol to ease the information burden for site owners. Along with standardization, there also needs to be information to encourage web owners to think about and set reasonable limits; a high crawl delay time may effectively ban web crawlers. There has been one related and successful development: the creation of HTML commands to robots that can be embedded in web pages. For example, the tag below, embedded in the head of a web page, instructs crawlers not to index the page (i.e., to discard the downloaded page) and not to follow any links from the page.

```
<META NAME=ROBOTS CONTENT="NOINDEX, NOFOLLOW">
```

The robots meta tag instructions are simple, extending only to a binary instruction to index a page or not and a binary instruction to follow links or not.

3.2.7 Critical review of existing guidelines

The robots.txt protocol only covers which pages a robot may download, but not other issues such as how fast it should go (e.g., pages per second). There are recognized, but completely informal, guidelines for this, written by the author of the robots.txt protocol (Koster, 1993). These guidelines aim to both protect web servers from being overloaded by crawlers and to minimize overall network traffic (i.e., the denial of service issue). Before building a crawl, programmers are encouraged to reflect upon whether one is really needed, and whether someone else has already completed a crawl and is willing to share their data. Data sharing is encouraged to minimize the need for crawling. If a crawler is really needed, then the robot should, amongst other things:

- Not crawl any site too quickly,
- Not revisit any site too often, and
- Be monitored to avoid getting stuck in 'spider traps': large collections of almost duplicate or meaningless pages.

3.2.7.1 Denial of service

The phrases 'too quickly' and 'too often', as noted above, invoke a notion of reasonableness in crawling speed. Decisions are likely to change over time as bandwidth and computing power expand. The figures mentioned in 1993 no longer seem reasonable: "Retrieving 1 document per minute is a lot better than one per second. One per 5 minutes is better still. Yes, your robot will take longer, but what's the rush, it's only a program" (Koster, 1993). University web sites can easily contain over a million pages, but at one page per minute they would take almost two years to crawl. A commercial search engine robot could perhaps save time by identifying separate servers within a single large site, crawling them separately and simultaneously, and also having a policy of updating crawling: only fetching pages that have changed (as flagged by the HyperText Transfer Protocol header) and employing a schedule that checks frequently updated pages more often than static ones (Chakrabarti, 2003; Arasu, Cho, Garcia-Molina, Paepcke, & Raghavan, 2001). In an attempt at flexible guidelines, Eichmann has recommended the following general advice for one type of web agent, "the pace and frequency of information acquisition should be appropriate for the capacity of the server and the network connections lying between the agent and that server" (Eichmann, 1994).

The denial of service issue is probably significantly less of a threat in the modern web, with its much greater computing power and bandwidth, as long as crawlers retrieve pages sequentially and there are not too many of them. For non-computer science research crawlers and personal crawlers, this is probably not an issue: network usage will be limited by bandwidth and processing power available to the source computer. Nevertheless, network constraints may still be an issue in developing nations, and perhaps also for web servers maintained by individuals through a simple modem. Hence, whilst the issue is less critical than before, there is a need to be sensitive to the potential network impacts of crawling.

3.2.7.2 Cost

The robots.txt protocol does not directly deal with cost, although its provisions for protection against denial of service allow site owners to save network costs by keeping all crawlers, or selected crawlers, out of their site. However, all crawling uses the resources of the target web server, and potentially incurs financial costs, as discussed above. Krogh (1996) has suggested a way forward that addresses the issue of crawlers using web servers without charge: to require web agents, including crawlers, to pay small amounts of money for the service of accessing information on a web server. This does not seem practical at the moment because micro-payments have not been adopted in other areas, so there is not an existing infrastructure for such transactions.

Commercial search engines like Google can justify the cost of crawling in terms of the benefits they give to web site owners through new visitors. There is perhaps also a wider need for crawler owners to consider how their results can be used to benefit the owners of the sites crawls as a method of giving them a direct benefit. For example, webometric researchers could offer a free site analysis based upon crawl data. There is a particular moral dilemma for crawler operators that do not 'give anything back' to web site owners, however. Examples where this problem may occur include both academic and commercial uses. There is probably a case for the 'public good' of some crawling, including most forms of research. Nevertheless, from a utilitarian standpoint it may be difficult to weigh up the crawling costs to one individual against this public good.

For crawler owners operating within a framework of social accountability, reducing crawling speeds to a level where complaints would be highly unlikely or likely to be muted would be a pragmatic strategy. The crawling strategy could be tailored to such things as the bandwidth costs of individual servers (e.g., don't crawl expensive servers, or crawl them more slowly). Sites could also be crawled slowly enough that the crawler activity would be sufficient to cause concern to web site owners monitoring their server log files (Nicholas, Huntington, & Williams, 2002).

3.2.7.3 Privacy

The robots.txt protocol and Koster's (1993) guidelines do not directly deal with privacy, but the robots.txt file can be used to exclude crawlers from a web site. This is insufficient for some privacy issues, however, such as undesired aggregation of data from a crawl, particularly in cases where the publishers of the information are not aware of the possibilities and the steps that they can take to keep crawlers away from their pages. Hence, an ethical approach should consider privacy implications without assuming that the owners of web pages will take decisions that are optimally in their own interest, such as cloaking their email addresses. Power relationships are also relevant in the sense that web site owners may be

coerced into passive participation in a research project through ignorance that the crawling of their site is taking place.

Some programmers have chosen to ignore the robots.txt file, including, presumably, those designed to collect email addresses for spam purposes. The most notable case so far has been that of Bidder's Edge, where legal recourse was necessary to prevent Bidder's Edge from crawling Ebay's web site (Fausett, 2002). This case is essentially a copyright issue, but shows that the law has already had an impact upon crawler practice, albeit in an isolated case, and despite the informal nature of the main robots.txt guidelines.

3.2.8 Guidelines for crawler owners

As the current protocol becomes more dated and crawler owners are forced to operate outside of the 1993 suggestions, researchers' crawlers could become marginalized; designed to behave ethically in a realistic sense but being grouped together with the unethical crawlers because both violate the letter of Koster's 1993 text. Hence new guidelines are needed to serve both to ensure that crawler operators consider all relevant issues and as a formal justification for breaches of the 1993 guidelines.

Since there is a wide range of different web hosting packages and constantly changing technological capabilities, a deontological list of absolute rights and wrongs would be quickly outdated, even if desirable. Utilitarianism, however, can provide the necessary framework to help researchers make judgments with regards to web crawling. It is important that decisions about crawl parameters are made on a site-by-site, crawl-by-crawl basis rather than with a blanket code of conduct. Web crawling involves a number of different participants whose needs will need to be estimated. This is likely to include not only the owner of the web site, but the hosting company, the crawler operator's institution, and users of the resulting data. An ethical crawler operator needs to minimize the potential disadvantages and raise the potential benefits, at least to the point where the net benefits are equal to that of any other course of action, although this is clearly hard to judge. The latter point is important, as introduced by Koster (1993): the need to use the best available option. For example a web crawl should not be used when a satisfactory alternative lower cost source of information is available.

The following list summarizes this discussion and builds upon Koster's (1993) recommendations. Note the emphasis in the guidelines on finding out the implications of individual crawls.

Whether to crawl and which sites to crawl

- Investigate alternative sources of information that could answer your needs, such as the Google API (www.google.com/apis/) and the Internet Archive (cost, denial of service).
- Consider the social desirability of the use made of the information gained. In particular consider the privacy implications of any information aggregation conducted from crawl data (privacy).
- For teaching or training purposes, do not crawl the web sites of others, unless necessary and justifiable (cost, denial of service).
- Be aware of the potential financial implications upon web site owners of the scale of crawling to be undertaken (cost).

    - Be aware of differing cost implications for crawling big and small sites, and between university and non-university web sites (cost).
    - Do not take advantage of naïve site owners who will not be able to identify the causes of bandwidth charges (cost).
    - Be prepared to recompense site owners for crawling costs, if requested (cost).
- Be aware of the potential network implications upon web site owners of the scale of crawling to be undertaken, and be particularly aware of differing implications for different types of site, based upon the bandwidth available to the target site web server (denial of service).
- Balance the costs and benefits of each web crawling project and ensure that social benefits are likely to outweigh costs (cost, denial of service, privacy).

Crawling policy

- Email webmasters of large sites to notify them that they are about to be crawled in order to allow an informed decision to opt out, if they so wish.
- Obey the robots.txt convention (Koster, 1994) (denial of service).
- Follow the robots guidelines (Koster, 1993), but re-evaluate the recommendations for crawling speed, as further discussed below (denial of service).

The above considerations seem appropriate for most types of crawling but create a problem for web competitive intelligence (Wormell, 2001; Vaughan, 2004a). Since an aim must be to gain an advantage over competitors (Underwood, 2001), an Adam Smith type of argument from economics about the benefits of competition to society might be needed for this special case, and if this kind of crawling becomes problematic, then some form of professional self-regulation (i.e., a social contract) or legal framework would be needed. This is a problem since there does not seem to be a professional organization that can fulfill such a need. Individual crawler operators also fall into a situation where currently only their internal moral code, perhaps informed by some computing knowledge, serves as an incentive to obey guidelines. From a social perspective, the pragmatic solution seems to be the same as that effectively adopted for spam email: formulate legislation if and when use rises to a level that threatens the effectiveness of the Internet.

For researchers, the guidelines can inform all crawler users directly, or can be implemented through existing formal ethical procedures, such as university ethics committees and, in computer science and other research fields, can be implemented through social pressure to conform to professional codes of conduct. Hopefully, they will support the socially responsible yet effective use of crawlers in research.

3.3 What can university-to-government web links reveal about university-government collaboration?

3.3.1 Introduction

This study analyses university-to-government web links, investigating whether the number of university-to-government web links can be used as an indicator of a university's research productivity; and what the reasons for the university-to-government link placement tell us about university-government relationships. Over recent years, there have been a number of acknowledged changes in the way contemporary research is carried out, with greater emphasis being placed on collaborative and government-driven, commercially-oriented research (Gibbons et al., 1994; Etzkowitz & Leydesdorff, 2000). In this context, webometric indicators may be useful to show collaborations that may not be reflected in traditional bibliometrics (Thelwall, 2004c).

This research focuses on the relationship between United Kingdom universities and government departments, as manifested by the web links from university web sites to government web sites. The reason for only focusing on one part of the Triple Helix model, as opposed to the helix as a whole, is so that it is feasible for the information to be gathered by a web crawler as opposed to relying on commercial search engines. The reason for choosing the university-to-government links was that there are a manageable number of easily identifiable university web sites, with those government sites also easily identifiable through their '.gov.uk' second/top level domain name. Looking for links to industry would be complicated by the vast array of top-level domain names that are potentially available to companies as they try to establish a more international feel or a particular image; it was also felt that links to government web sites were less likely to be frivolous.

Looking at universities in the United Kingdom also allows for the number of links to be compared to an acknowledged indicator of research productivity, the Research Assessment Exercise (RAE, 2005).

3.3.2 Methodology

The investigation was carried out in three parts: first, establishing whether there was a correlation between the number of university-to-government web links and a university's research success; second, establishing whether this correlation was stronger than a correlation between the total outlinks from a web site and a university's research success; and finally, a classification of a sample of the university-to-government web links to establish the reasons for link placement. This is broadly in line with a recommended link analysis methodology (Thelwall, 2004a).

The data set was taken from a publicly available database containing the link structure of 125 UK universities (http://cybermetrics.wlv.ac.uk/database/). This database is based on a web crawler designed for the collection of data for the academic community (Thelwall, 2003a).

3.3.2.1 Establishing a university's research quality

A figure to express the research quality of a university was calculated based on the results of the 2001 Research Assessment Exercise (i.e., the most recent one). These scores were taken to be on a linear scale, which is the standard interpretation (Education Guardian, 2001) and has been used previously in academic research (Thelwall 2002e; 2004a). These scores were then

multiplied by the number of researchers in the department, totalling the products for all the departments in a university. This total was then divided by the number of full time equivalent academic staff (Education Guardian, 2001; Thelwall & Harries, 2004b). Such normalisation is essential because UK universities have greatly differing sizes: correlations for data without normalisation could be explained by size differences.

A program calculated the number of university-to-government web links for each university. This figure was normalised by dividing by the full time equivalent number of staff, and then compared to the university's research quality figure using Spearman's rank correlation test. This procedure was repeated for the total number of outlinks for each university.

3.3.2.2 Classification of reasons for hyperlinks

A classification of the hyperlinks was carried out using content analysis techniques, (Krippendorff, 2004). Following Wilkinson, Harries et al. (2003), a random selection of ten web links was extracted from each of the 125 universities' lists of web links to create a single file of random university-to-government web links that would not be heavily biased towards larger universities. For those universities where there were not ten links to government web sites, as many links as were available were added to the file. This file was then randomly sorted, and the first 400 were used as part of the classification exercise.

Although classifications of university hyperlinks have been carried out previously (Wilkinson, Harries et al., 2003; Bar-Ilan, 2004a), the classification schemes reflect the specific research questions addressed, and as such are inappropriate for the classification of university-to-government web links. The classification initially used three broad headings, links were attributed as representing an arbitrary relationship; an existing research based relationship; or a desire for there to be a relationship. Further breakdowns in the classification system were integrated in an inductive system, with finer grained classification being integrated as each of the pages were visited. When a link covered more than one of the categories, it was classified under the section that represented the highest level of research relationship that the page included.

Classification of links was carried out by examining the source web page to establish the link creation context. If necessary the context of the page was determined by looking at whereabouts the web page fitted into the university's web site. For any pages where it was not possible to classify the link by looking at the current page due to either the web page or the link no longer being present, a cached copy of the page was retrieved from the Internet Archive's Wayback machine (Internet Archive, 2005). To reduce classifier bias more than one person should carry out any classification exercise, but a single classifier is acceptable for exploratory research (Thelwall, 2004a).

3.3.3 Results

Spearman's rank correlation coefficient for the correlation between the number of university-to-government web links per full time equivalent member of staff and the research productivity per full time equivalent member of staff gave a value of .168 (n = 125); which is not statistically significant. Spearman's rank correlation coefficient for the total number of outlinks was .738, which is statistically significant at the 1% level. Figure 3-2 and Figure 3-3 express the correlations in the form of a scatter graph. Whilst the correlation is clearly visible in Figure 3-3, none is apparent in Figure 3-2.

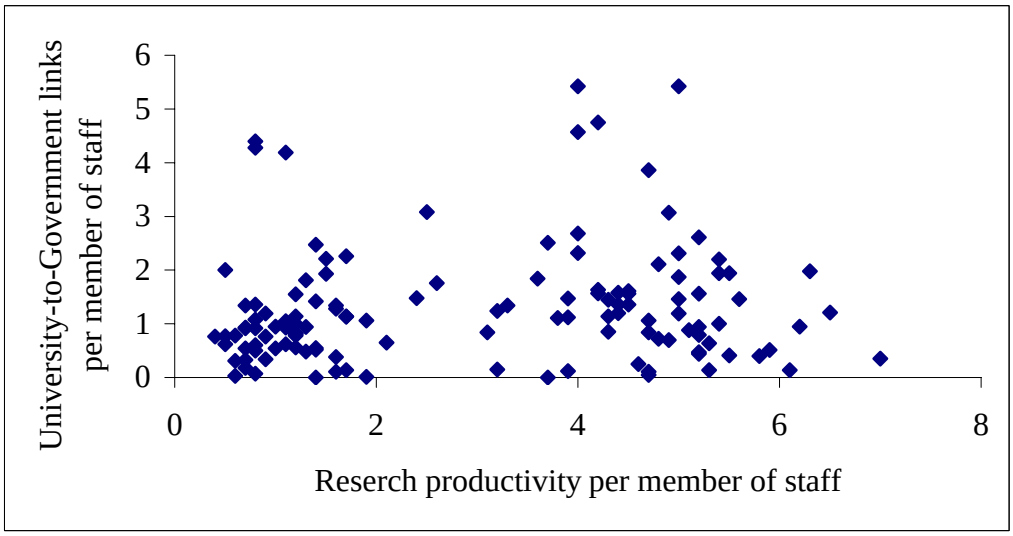

Figure 3-2 Scatter graph comparing the number of university-to-government links with research productivity per member of staff

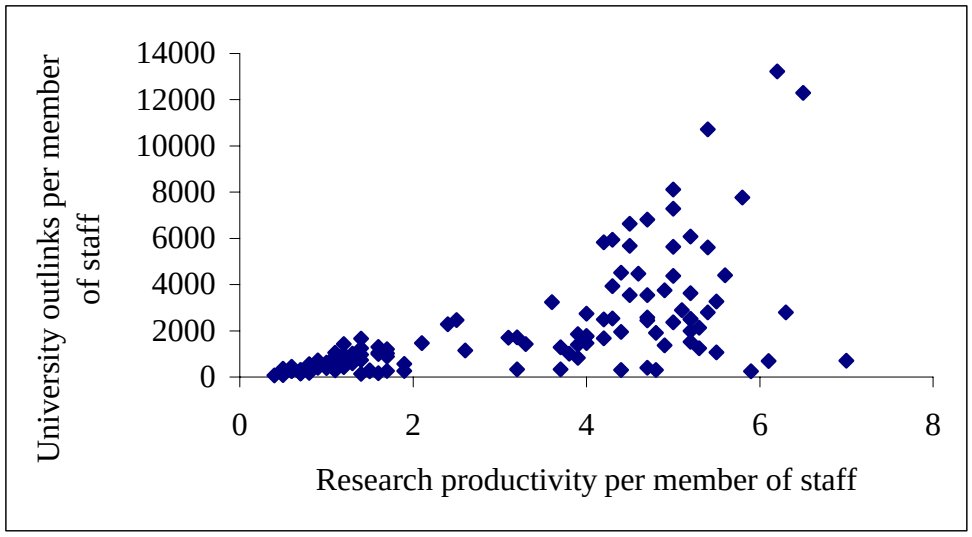

Figure 3-3 Scatter graph comparing total university outlinks and research productivity per member of staff

Table 3-3 shows the final classification scheme. For those links that could not be found either on the internet or on the Internet Archive, arguably the transitory nature of the link means that the relationship is unlikely to be permanent.

Table 3-3 Classification scheme for web links including the number of links in each category

Relationship inferred from web links	Number of links
1. Expressing a research relationship	
Previous work	2
A client	8
Funding / Sponsorship	12
Collaboration	13
Undisclosed	7
2.Expressing desire for relationship:	
Careers	16
Funding information	9
Other	2
3.Expressing an arbitrary relationship	
Academic resource	190
References	37
Other information	84
Link placed in error	2
4. Unable to determine relationship	18

3.3.4 Discussion

The results show a strong correlation between a university's research productivity and its total outlinks. Interestingly, the university's number of university-to-government links does not show a better correlation, and is not even statistically significant. There are a number of possible reasons for this:

- A greater percentage of university-to-non-government links reflect a research relationship than university-to-government links.
- University web sites don't generally indicate the governmental origin of their research funding.
- The university-government relationship is not as important as other relationships to a university's research productivity.
- There were too few links for a pattern to emerge.

In fact, most UK government funding is indirect, via funding councils. The funding councils typically have .ac.uk academic domain names. The government's role is the provision of the funding, and the overall direction in which the funding should be allocated. Hence the lack of university-government links may reflect that the government does not actually exist as a serious part of the research funding process: the key players are the academics who decide upon the allocation of funds (Whitley, 2000). In other words, the university-government relationship is actually a university-university relationship in many ways, particularly from the perspective of individual researchers.

3.3.5 Conclusion

This research shows that whilst there is a correlation between a university's research productivity and the total number of outlinks, there is not a strong correlation between a

university's research productivity and the number of university-to-government links. A classification of the links suggests that the reason for this is that the links are placed predominantly to highlight information resources and not real collaborative efforts.

The obscuring of the university-government connection has also been discussed as a possible cause of the results, with the government hidden behind research councils, which themselves rely heavily upon academics for their decisions.

It is likely that between each pair of top-level domains there are an identifiable set of reasons for the placing of web links, and these reasons differ in proportion for each. Further research is needed to identify these differences so that webometrics can be used to provide indicators to more specific questions.

3.4 Academic web links and collaboration

3.4.1 Introduction

This investigation builds upon the above-mentioned study (Stuart & Thelwall, 2005), by seeking to discover why university links target different types of domain (besides *gov.uk*). An extended classification approach is used and incorporates an analysis of the source page of the web links. Both the target domain and the source page owner are likely to affect the range of reasons why a link is placed on a web page. University web sites host a variety of web pages from different people and the objectives of the web page author are likely to affect the reasons links are placed on a web site. For example, the majority of links on a library page are probably placed to highlight information resources, whereas the links on an academic's homepage are likely to be there for an eclectic mix of reasons. Organisations from different sectors of the web are likely to provide different content on their web sites; where sites provide useful content it is likely they will be linked to, regardless of whether or not there is a relationship between the organisations, whereas if the site does not contain a high standard of content, they are unlikely to be linked to unless there is a collaboration. As it has been suggested that the research councils play an important role in mediating between the UK government and universities, an analysis has also been carried out on the reasons web sites link to each UK research council.

The research questions addressed are:

- Does a significant proportion of web links from UK university web sites reflect a collaborative relationship between the university and the target organisation?
- Are there differences in the proportion of links that reflect a collaborative relationship depending on the second/top level domain of the target organisation?
- Does the propensity for web links to reflect collaboration depend on the source page owner type?

3.4.2 Methodology

The investigation was carried out in three parts: collecting a sample of web links; a qualitative analysis of the web links using content analysis techniques (Krippendorff, 2004); and a quantitative analysis of the results of the qualitative data.

3.4.2.1 Data collection

The data set was taken once again from a publicly available database containing the link structure of 125 UK universities in June 2004 (freely available from http://cybermetrics.wlv.ac.uk/database/) (Thelwall, 2003a). It would have been possible to retrieve a selection of URLs using the advance search interface of certain search engines, such as Google and AltaVista, but these were not used, as it is not always clear as to when the pages were indexed or how deep the search engine crawlers have gone. Hence the more controlled crawler sample is preferred.

A program was written to extract from this database the URLs of university web pages that linked to government and industry web sites. Government and industry web sites were determined by their second/top-level domain names. Of the twelve second-level domain names provided under the UK top level domain name (Nominet, 2006) pages that linked to five were examined: gov.uk, co.uk, org.uk, plc.uk, ltd.uk. Pages that linked to a research council web site were also extracted, based upon previous work suggesting that government to university relationships may be hidden by the research councils which are in the ac.uk domain. The research councils were identified from the Research Councils web site, and also included was the Art & Humanities Research Board, which later became a research council in its own right (Research Councils UK, 2006).

For each of the domains under investigation a selection of 10 URLs was sampled randomly (with replacement) and used to create a combined list. By taking only 10 URLs from each of the universities the links were not heavily biased in favour of the larger universities. The combined lists were then randomised and the first 200 were taken from each list and used as a sample. This came to a total of 2,600 URLs that were classified by the first classifier, ten percent of which were classified by the second classifier. As web pages are known to vary over time each of the web pages were downloaded on the same day, and saved onto CD ROM so that the researchers could be sure of classifying the same page (Wilkinson, Harries et al., 2003).

3.4.2.2 Link and source page classification

In order to answer the research questions, both ‘source page owner’ and ‘reason for link placement’ were investigated for the random sample. To determine the type of relationship appearing on the web it is necessary not only to establish why the links are being placed, but also who are placing the links; if all links that expressed a collaborative relationship were being placed by individuals it would not necessarily indicate a formal relationship. Other potential facets such as link type and target page type were rejected as although they could provide interesting information, they are not necessary in determining the existence of collaborative relationships on the web, and would have added considerably to the research time needed due to the size of the sample.

Some previous webometric studies have included a classification of web links (e.g., Kim, 2000; Bar-Ilan, 2004a), but as these schemes were established to address different specific questions, they were unsuitable for this research. It has been suggested that the wide diversity in web uses and enquiries means that there is unlikely to be a widely agreed typology (Harries et al., 2004), a point that is supported by the variety of schemes proposed for the relatively well established field of citation analysis.

For this study the classification scheme was established by post coordinate clustering: visiting a random selection of the web links and then devising appropriate categories for links

that appear to be similar (Thelwall & Harries, 2004a). An iterative classification scheme based on the whole sample (Thelwall, 2001d) was not used as it is not appropriate for more than one classifier, and more than one was used to enhance reliability. The post coordinate clustering utilised a preliminary classification scheme, which was itself based on the research question and previous studies, and a random selection of 10% of the sample URLs from each of the domains.

Although there are a number of different facets that could be examined (Bar-Ilan, 2004a), it was decided that only two would be looked at:

- Source page owner – The preliminary classification scheme is based upon two previous classifications of university web pages (Bar-Ilan, 2004a; Harries et al., 2004) department or school page; institute or research group page; non-academic/non-research unit page; individual's page; and resource list. The inclusion of the 'non-academic/research unit' classification reflects the notion of hybrid organisations that may appear on university web pages.
- Reason for link placement – This is the facet that most closely reflects the perspective of the research question. Does the link placement reflect a collaborative relationship between the organisations, and if so what is the nature of that relationship? As it was expected that the final classification could potentially include a high number of reasons for link placement, with subtle differences, it was decided that nested categories would be used as it would be difficult to get inter-classifier agreement on the finer levels. The preliminary classification was based on Wilkinson, Harries et al.'s (2003) classification scheme for web links. These categories were divided into the broad headings of 'Reflecting a relationship' (research relationship; business relationship; page creator or sponsor; and other relationship) and 'Arbitrary relationship' (research support; student learning material; information for students; tourist information; recreational; libraries and e-journals).

A classification protocol was established for the 'reason for link placement' to improve inter-classifier reliability, as this was considered the most important aspect of the study (see Appendix 1). Whilst the protocol gives a detailed breakdown of the collaborative relationship categories, the difficulty in devising a protocol for non-collaborative relationships was thought to outweigh any benefits gained. The more detailed classification provided in the results section is merely for illustrative purposes.

The first researcher classified all the links, and a second researcher classified 10% of the pages as a reliability test.

3.4.3 Testing for statistical significance

Applying chi-square tests to a set of hypotheses tested the statistical significance of the results. Hypotheses are important so that specific instances of statistical significance could be investigated, as opposed to just having the statistical significance of the table as a whole. A subset of all possible research hypotheses was chosen to control error levels, which can be problematic when a large number of hypotheses are simultaneously tested (Wilcox, 2003).

3.4.3.1 Target domains

It is expected that there would be a significant variation in link placement reasons for different target domains between:

- *Research councils and various UK second-level domains (excluding ac.uk)*: it would be expected that more of the links to research council web sites reflect collaboration than those to the second-level domains;
- *Research councils*: different fields utilise the web in different ways (Harries et al., 2004);
- *Second-level domains*: the different second level domains host different sorts of sites with differing collaboration likelihoods;
- *co.uk and gov.uk*: large numbers of non-collaborative relationship links to a few sites e.g., amazon.co.uk, are likely to swamp those collaborative relationship links;
- *A research council (ESRC) and both co.uk and gov.uk*: it has been suggested previously that part of the reason for proportion of collaborative links between university web sites and government web sites is due to these links being hidden through the research councils (Stuart & Thelwall, 2005).

The Economic and Social Research Council was used as an example of a research council as social science linking behaviour lies somewhere between the extremes of the hard sciences and the arts and humanities (Tang & Thelwall, forthcoming).

3.4.3.2 Source page owner

It is expected that there would be a significant variation in the link placement reasons of different source page owners for:

- 'Library & information service pages' and both 'Externally focused university organisation pages' and 'Department or school pages': the library pages are likely to host many more links for non-collaborative relationship reasons;
- 'Department or school pages' and 'Research group or institute pages': as well as providing information about the department, department web pages are likely to also host a number of links to a number of web resources for staff and students, as such there are more likely to be links reflecting collaborative relationships from research group or institute pages.

It is unlikely that there will be a statistical difference in the proportion of links reflecting a collaborative relationship between an affiliated independent page and an individual's page, both are likely to consist of an eclectic mix of resources, references, and those that reflect a collaborative relationship.

3.4.4 Results

3.4.4.1 Source page owner classification scheme

Table 3-4 shows the final classification scheme for the 'Source page owner' facet. Establishing the page type was not always possible by looking at the page alone; more information was gleaned by visiting adjacent university pages that either had links on the source page or could be found by truncating the source page's URL.

Table 3-4 Classification scheme for university web pages

Source Page Owner	Examples
1.Department or school page	Departmental links list; module list; homepage
2.Research group or institute page	Home page, link list of an identifiable research group or institute
3.Service department page	Human resources; graduate office; main university pages
4.Externally focused university organisation page	Electronic journals; conferences; learned societies
5.Affiliated independent group page	Student group/recreational group/ student union
6.Individual's page	Academic homepage, student homepage
7.Library & information service page	Catalogue, page of information about the library
8.Careers service page	Jobs lists, list of potential recruiters

Whilst certain of the categories appear self-explanatory e.g., 'Department or school page', or 'Library & information service page'; others need further explanation. 'Service department page' was determined as a department whose function was to facilitate the running of the university, e.g., a human resource department or a graduate office. Although both 'Library & information service page' and 'Careers service page' can be seen as falling within this category, they were counted separately due to the large number of outlinks that they were expected to contain for non-collaborative reasons.

In addition to the departments that deal with the running of the university and its teaching and research activities, numerous other groups and organizations are hosted on a university's web site. These were split into two categories: 'Externally focused university organisation' was used to describe a page which was designed for external as well as internal use, e.g., electronic journals, conferences, learned societies; 'Affiliated independent group' was used to describe groups that were internally focused, but weren't part of the official university organization, e.g., recreational group, student union group.

3.4.4.2 Target page classification scheme

Table 3-5 shows the classifications for link placement reasons. As well as the two broad categories of 'Reflecting a collaborative relationship' and 'Non-collaborative relationship', there were also a number of links that couldn't be classified due to either the page no longer being available, or the page having changed since the crawl had taken place and the link no longer being on the page. As mentioned above, due to the difficulties in classifying non-collaborative relationships, and their irrelevance to the research question, detailed

classification is only provided for illustrative purposes. See appendix 1 for the full classification protocol.

Table 3-5 Classification scheme for university outlinks

Reason for Link	Description
1.Reflecting a collaborative relationship (formal or informal)	
a. Research relationship/collaboration	Link placed to a target because they have a joint research project with the source.
b. Business client	The source has carried out work on behalf of the target.
c. Previous employer	The source page owner previously worked for the target owner.
d. Sponsorship/Funding	Target owner has provided sponsorship or funding for a source page owner activity, e.g., a conference or research.
e. Shared interests	Organisations working together towards mutual goals, e.g., making sure courses provide the required skills
f. Reflecting membership of an organisation	e.g., link on a homepage to professional organisation or on a department page to show accreditation.
g. Informal relationship	e.g., somebody speaking at a conference
h. Contractor	Target page owner has provided work for source page owner.
i. Undisclosed relationship	e.g., 'supported by' or logo of target page, but no indication of why.
2. Non-collaborative relationship	
a. Information resource	e.g., an e-journal; a site containing statistical information
b. Information reference	e.g., link to the government data protection act as a reference in a university's computer policy
c. Acknowledgement	e.g., thanks for the right to publish an article
d. Research support	e.g., information on how to get research funding
e. Welfare information	e.g., workers' rights; visas; student funding
f. Tourist information	e.g., addition information for those visiting the university
g. Organisation of interest	e.g., a link on a careers web site where the target organisation is of primary interest rather than the information on the web site
h. Product link	e.g., where to download adobe acrobat
i. Recreational	e.g., a person's favourite football team
j. Placed in error	e.g., a link to an unrelated site on a site map
3. Unable to determine reason	
a. Link no longer on the page	
b. Page no longer available	

3.4.4.3 Inter-classifier consistency

The first classifier classified 2,600 links and pages, out of which a second classifier classified 260 links and pages. Classifier reliability was determined by calculating Krippendorff's alpha coefficient (Krippendorff, 2004); this takes into account the chance agreement, the magnitude of the misses, and adjusts for whether the variable is nominal, ordinal, interval, or ratio (Neuendorf, 2002).

Of the 213 source pages that were not cross-classified as 'unable to determine reason', 67 were classified with different page types; Krippendorff's alpha was 0.6319. No two of the classifications were consistently used interchangeably, although 13 discrepancies were caused by one of the classifiers choosing 'School page' and the other choosing 'Service department page', and 13 of the discrepancies were caused by one of the classifiers labelling a page as a 'Service department page' and the other labelling it an 'Individual's page'. The overall frequency of source page types were similar for both of the classifiers with the exception of 'Externally funded pages', which the first classifier attributed to 24 of the 213 pages in comparison to the second classifier's eight pages, and 'Individual's page' which the second classifier attributed to 31 pages in comparison to the first classifier's 19 pages. There were only three cases where one classifier assigned 'Externally focused' and the other classifier assigned 'Individual's page'.

Even though a classification protocol was used (see appendix 1) Krippendorff's alpha coefficient for classifier agreement for link placement was only 0.6749. The level of reliability rose to 0.7995 at the coarser level of whether or not the link reflects a collaborative relationship or a non-collaborative relationship. As there is a significant increase in Krippendorff's alpha for the coarser level of classification these results are discussed in this paper.

3.4.4.4 Do more links to some domains reflect a non-collaborative relationship?

Table 3-6 shows the reasons for link placement according to the target page domain for those pages that could be classified.

Multiple comparisons on a single table necessitate the Bonferroni correction (Wilcox, 2003, p.443): for the results of n tests carried out on a single table to be statistically significant at the 5% level it is necessary for $p<0.05/n$.

The chi-square tests show that the differences in the reasons for link placement are statistically significant at the 1% level, when comparing:

- all the domains (chi-sq=143.9, d.f.=12, $p<0.001$);
- the research councils (chi-sq=97.35, d.f.=7, $p<0.001$);
- between esrc.ac.uk and co.uk (chi-sq=36.24, d.f.=1, $p<0.001$);
- between esrc.ac.uk and gov.uk (chi-sq=18.43, d.f.=1, $p<0.001$).

The differences were statistically significant at the 5% level when comparing:

- the five top level domains (chi-sq=17.14, d.f.=4, $p<0.005$).

The differences in the linking between gov.uk and co.uk were not found to be statistically significant (chi-sq=3.055, d.f.=1, $p<0.1$).

Table 3-6 Reason for link placement - by target page

Target page domain	Domain use	Reason for link placement	
		Collaborative relationship	**Non-collaborative relationship**
co.uk	Nominally commercial enterprises, but could be any	**14** (9%)	**142**
org.uk	Nominally non-commercial organisations, but could be any	**30** (21%)	**112**
ltd.uk	Registered companies	**29** (19%)	**126**
plc.uk	Registered companies	**37** (27%)	**102**
gov.uk	Government bodies	**22** (16%)	**119**
ahrb.ac.uk	Arts & Humanities Research Board	**37** (25%)	**113**
bbsrc.ac.uk	Biotechnology and Biological Sciences Research Council	**42** (26%)	**122**
cclrc.ac.uk	Council for the Central Laboratory of the Research Councils	**37** (22%)	**129**
epsrc.ac.uk	Engineering and Physical Sciences Research Council	**90** (55%)	**73**
esrc.ac.uk	Economic and Social Research Council	**60** (38%)	**99**
mrc.ac.uk	Medical Research Council	**23** (15%)	**130**
nerc.ac.uk	Natural Environment Research Council	**36** (23%)	**122**
pparc.ac.uk	Particle Physics and Astronomy Research Council	**24** (16%)	**125**

3.4.4.5 Do more links from certain types of source pages reflect a non-collaborative relationship?

Table 3-7 shows the reason for link placement by the type of source page.

The chi-square tests show that the differences in the reasons for link placement are statistically significant when comparing:

- All the source pages (chi-sq= 396.4, d.f.=7, $p<0.001$);
- Library & information service pages and externally focused university organisation pages (chi-sq=147.9, d.f.=1, $p<0.001$);
- Library & information service pages and school pages (chi-sq=56.30, d.f.=1, $p<0.001$);
- Department or school pages and research group or institute pages (chi-sq=78.05, d.f.=1, $p<0.001$).

There is not a statistically significant difference between an affiliated independent group's page and an individual's page (chi-sq=0.3337, d.f.=1, p<1).

Table 3-7 Reason for link placement - by source page owner

Source page owner	Reason for link placement	
	Collaborative Relationship	**Non-collaborative Relationship**
Department or school page	75 (22%)	273
Research group or institute page	195 (54%)	168
Service department page	42 (10%)	362
Externally focused university organisation page	84 (49%)	86
Affiliated independent group page	14 (30%)	33
Individual's page	68 (26%)	196
Library & information service page	2 (1%)	250
Careers service page	1 (1%)	146

3.4.4.6 Estimated number of collaborative links

Table 3-8 shows the estimated number of collaborative links from universities to other domains, at the time of the crawl. The table highlights the fact that a high proportion of links placed for collaborative reasons does not necessarily equate to a high number of collaborative links, or vice versa. This is exemplified by the high number of collaborative links to the co.uk domain, second only to the number of collaborative links with the org.uk domain, despite the extremely low number proportion of collaborative links.

Table 3-8 Estimated number of collaborative links

Target domain	**Number of links**	**Proportion of links that are collaborative**	**Estimated number of collaborative links**
.co.uk	415,929	9%	37,327
.org.uk	200,453	21%	42,349
.ltd.uk	309	19%	58
.plc.uk	138	27%	37
.gov.uk	130,092	16%	20,298
.ahrb.ac.uk	2,780	25%	686
.bbsrc.ac.uk	2,827	26%	724
.cclrc.ac.uk	276	22%	62
.epsrc.ac.uk	6,109	55%	3,373
.esrc.ac.uk	6,660	38%	2,513
.mrc.ac.uk	4,467	15%	672
.nerc.ac.uk	4,661	23%	1,062
.pparc.ac.uk	1,328	16%	214

3.4.4.7 Significance of the results

Whilst the results of the chi-square tests mainly confirm the hypotheses, there are a number of anomalies:

- The difference between the five second/top-level domains was not as statistically significant as expected.
- There was not a statistically significant difference in the linking between co.uk and gov.uk.
- There was a wide range of differences amongst the research councils, and it was not the case for all of them that they had a higher proportion of collaborative links than the second/top-level domains.

The lack of a more significant difference between the five second/top-level domains is likely to be due to the wide variety of web pages that are hosted within each of the domains, and the same types of pages appearing over different domains. The variety of information that is available within each of the domains makes it difficult to attribute reasons for differences in the propensity for links to represent a collaborative relationship. The previous study into the number of links between universities and government reflecting a research relationship suggested that the low number might have been due to these relationships being hidden via the research councils. Whilst this study shows that generally more links to research councils reflect a collaborative relationship than links to government web sites, this is not always the case. There is a lower proportion of collaborative links to the Medical Research Council, and the same proportion to the Particle Physics and Astronomy Research Council.

That there is a wide diversity in the reasons for linking to different web sites is likely to be due to a combination of both the linking practice of the different research communities the sites are directed towards, and the information that is provided by the sites themselves. The differences in who is linking to these sites are shown in a comparison of who is linking to the Engineering and Physical Sciences Research Councils and who is linking to the Medical Research Council. There is a significantly higher proportion of links to epsrc.ac.uk from 'Service departments' and 'Individual's pages', which are more likely to reflect a collaborative relationship; and a significantly lower proportion of links to 'Library and information service pages', which are more likely to reflect a non-collaborative relationship.

Table 3-9 Source pages that are linking to EPSRC and MRC

Source Page Owner	EPSRC	MRC
Department or school page	34 (21%)	23 (15%)
Research group or institute page	57 (35%)	34 (22%)
Service department page	29 (18%)	46 (30%)
Externally focused university organisation page	10 (6%)	1 (1%)
Affiliated independent group page	0	0
Individual's page	27 (17%)	15 (10%)
Library & information service page	2 (1%)	30 (20%)
Career's page	4 (3%)	4 (3%)

The differences in the reasons for link placement depending on the source page were found to be statistically significant, and follow the expected order: those whose job is predominantly providing external information, e.g., libraries and careers services, have a relatively low number of collaborative links in comparison to non-collaborative links; those who would be expected to encourage collaborative relationships, e.g., research groups, have a relatively high

number of collaborative links; and the pages which provide information about themselves and provide access to external information, e.g., school departments and an individual's page, fall somewhere between the extremes.

3.4.5 Discussion

The low proportion of links that reflect a collaborative relationship causes difficulties in the use of web links in providing an indicator of collaboration between organisations. It is clear from the results of this study that it is difficult for an assertion to be made with any degree of confidence regarding the nature of the relationship between a university and another organisation based purely on the fact there is a web link from one to the other. However, that a significant proportion, if not a majority, of university outlinks reflect collaborative relationships means the link structure of the web has potential as a source of organisation collaboration indicators if appropriate methods are adapted for the filtering out of non-collaborative links.

The low proportion of links that reflect a collaborative relationship for both of the facets examined in this paper, i.e., source page type and target domain, means that it is likely that a number of facets will need to be used in conjunction with one another if assertions are likely to be made with a high degree of certainty. As well as source page and target domain such facets may include: the number of links per page; the position of the link on the page; the content of the anchor, or near anchor, text; target page type; and the page owner. Whilst the automatic assigning of information such as target domain is relatively simple to add to a web crawler, determining more subjective facets such as page type or field type is much more difficult. How can a web crawler be programmed to determine such facets when human classifiers cannot reach a high level of agreement?

The problems of automatically extracting collaborative links are emphasised by the university to co.uk links. Although only nine percent of the links represented a collaborative relationship, the massive number of university to co.uk links meant that this was a sizeable number of all collaborative links and should not be ignored.

Problems of inter-classifier consistency are not uncommon when classifying web pages and have been found in previous studies (e.g. Harries et al., 2004; Haas & Grams, 1998). The difficulties tend to depend on the level of inference that is necessary on the part of the classifier (Haas & Grams, 1998). For some facets inference is not necessary, for example the language of the page and the placement of the link can reach inter-classifier consistency of 100% (Bar-Ilan, 2005b), whilst for others such as reason for link placement inference is usually necessary unless the page author has been explicit in the reasons for link placement (Haas & Grams, 1998). The web is also filled with numerous pages which seem to defy classification: hybrid pages, where an identifiable individual deals with the web presence on behalf of a larger organisation; half formed pages which have been started before being left to rot; pages that were once part of a recognisable entity but have all but been orphaned; and pages that are incomprehensible merely due to the whim of the author. These problems may then be exacerbated with the propensity of people to regularly utilise terms such as homepage, or department page, without a concrete notion of what they mean. Problems with the classification of certain facets are likely to increase as new information is put on the web without the removal of old information.

The work necessary for the automatic classification of whether web links represent a collaborative relationship is unlikely to produce a high enough success rate, and as such would

be pointless at this time. As improvements are made in terms of automatic content analysis it may be worth returning at a future date.

The web has the potential to provide a richer source of information regarding the relationships between different individuals, organisations, and countries than traditional bibliometric indicators. The richness is in part due to the web's unstructured nature, relationships do not need to be pigeonholed, or placed for a specific reason; they can be placed for any reason. This same richness also means that it is much more complicated to extract collaborative information than through looking at co-authorship of papers or co-inventorship. Gathering collaborative information from the web is analogous to collecting it from citations with similar difficulties in separating the wheat from the chaff. In the same way that it only becomes clear that citations reflect relationships when the relationships are known, it is likely that there are many relationships that will not have been identified within this study, as there was not enough information provided on the web site to make those assertions.

Gathering information about collaborative relationships between organisations from web links would need a lot of human intervention, meaning that it is not a financial avenue for a large-scale study. As web sites would need to be investigated individually to determine whether or not the web links do reflect collaboration, the initial investigation into the link structure would become superfluous.

This study has a number of limitations. The time lag between the crawl and the classification may have had some affect on the results; it is possible that pages reflecting collaborative relationships are more volatile than non-collaborative relationships and as such a higher proportion of those links that could not be classified reflected a collaborative relationship. It is also possible that the exclusion of non-dynamic pages affects the proportion of non-collaborative links, it is likely that many library and information service pages are dynamically generated, however, inclusion of such pages would probably result in the same links appearing on numerous dynamically generated pages.

3.4.6 Conclusion

This study shows that a significant number of the links that are placed on university web sites to other organisations reflect a collaborative relationship between the different web page owners, and these links are therefore a rich source of information regarding collaborative relationships. The study also shows that there are statistically significant differences in the proportion of links that reflect a collaborative relationship depending on both the source page owner and the target page top-level/second-level domain. However, the low proportion of links that reflect a collaborative relationship means that much more research is necessary before web links can be used as a reliable indicator of collaboration between organisations; merely the existence of a link between two organisations is not a good enough indicator of a relationship between the two organisations.

Future investigations into extracting collaborative information from the web will need to investigate the linking practices on different domains, and in different countries, as these are likely to contain differing reasons for link placement. A potential future area of investigation is into those sites where there are a number of web links between two organisations, analogous to White's re-citations (2001). Whilst single web links have not been shown to provide an indicator of a collaborative relationship between two organisations, this may change for clusters of links.

3.5 Investigating Triple Helix relationships using URL citations: A case study of the UK West Midlands automobile industry

3.5.1 Introduction

This exploratory study investigates the potential use of URL citations as weak benchmarking indicators to estimate levels of collaboration between organisations within the different sectors of the Triple Helix model: university, industry, and government. Weak benchmarking indicators are indicators that may provide useful information, but have not met the required reliability and validity criteria to be used in an evaluative role (Thelwall, 2004c). In other words their values are indicative or suggestive rather than definitive. One useful application is in identifying outliers: web sites that do not fit the expected pattern for their type.

This initial exploratory study into university-industry-government relations focuses on one specific industry in a small area of the UK, the automobile industry in the West Midlands, to addresses the question: Can URL citations be used as weak benchmarking indicators to determine levels of collaboration between organisations within the different sectors of the Triple Helix model? This is a case study style of research, which is appropriate for a new research topic.

3.5.2 Research methodology

Car production is an established industry in the West Midlands and by focusing on the more significant companies in the automobile industry, the region and industry provides a number of web sites that is not too large for the capabilities of the data collection methods adopted. The West Midlands region incorporates thirteen higher education institutions (West Midlands Higher Education Association, 2005), thirty nine principal local authorities (DirectGov, 2005), and the automotive industry unit of the Department of Trade and Industry highlights fourteen organisations on their web site that play a significant role in the automobile industry in the region (DTI, 2005). For convenience, all higher education institutions in the study will be described as universities, even though most do not hold this legal status (e.g., not having the power to award PhDs).

Organisational home pages can have multiple URLs and some organisations use more than one domain name. For example, Wolverhampton University utilises two domains:

.wlv.ac.uk
.wolverhampton.ac.uk

It is also the case that some small organisations use space on another organisation's server, and do not have a separate domain name. Information on the different domain names for universities is readily available (University of Wolverhampton, 2005) and, where applicable, multiple domain names were used as search terms. Each of the industry and local government web sites were also visited to determine whether the identified domain was still the principal focus. When a page was redirected to another URL both the old and the new URLs were used as search terms.

The Google API allows for the automatic retrieval of up to 1,000 queries per day, and queries may be sent to the Google API database in a number of different programming languages. It allows for the retrieval of a range of information for each of the queries,

including the estimated total number of hits for a particular query, and the retrieval of the URLs for a limited number of the hits (the first 1000).

3.5.2.1 Number of pages indexed

As Google does not crawl and index every web page of every web site it is necessary to determine that each of the web sites involved in this investigation have had sufficient pages crawled and indexed to allow for conclusions to be drawn about URL citations placed on the sites. A program was written to send queries to Google API, utilising the Google's 'allinurl:' facility, for example:

allinurl:.wlv.ac.uk

This query would list indexed pages with '.wlv.ac.uk' in their URL. The 'allinurl:' facility was used as opposed to 'site:' as the functionality of the 'site:' facility was found to often be inaccurate unless used in conjunction with some other keyword; it often states that a query produced no hits despite the addition of a more restrictive keyword producing hits.

3.5.2.2 Number of URL citations between web sites

Another program was written to determine the number of URL citations between the different organisations. Although Google allows the retrieval of pages that link to a specified page, it does not allow for the retrieval of pages that link to a specified web site, so it is not as appropriate as searching for URL citations, mentions of a URL in a web page whether hyperlinked or not (Kousha & Thelwall, 2005). This was accomplished by utilising the Google Web API's phrase search and site restricted search capabilities, for example:

".wolverhampton.gov.uk" site:.wlv.ac.uk

This query would retrieve information from the University of Wolverhampton domain that included .wolverhampton.gov.uk, the domain name of the Wolverhampton local government web site. The appearance of an organisation's URL in a web page is likely to often be a link to that organisation's web site.

3.5.2.3 Confirmatory URL citation analysis

In order to interpret URL citation counts the link analysis methodology uses a classification of the purposes of web links (Thelwall, 2004a). The Google API limits the URLs that can be received to any ten consecutive hits in the first thousand returned. Whilst it would be possible to retrieve a random sample of non-consecutive URLs from the first thousand hits, the retrieval would mean repetition of similar queries, and as the number of automatic queries that can be sent in one day are restricted, it would mean a lengthy data collection exercise. Even selecting ten consecutive URLs from a random position in the first thousand hits would double the time required; an initial query would be needed to establish how many URL citations there were before a random group of ten could be extracted on a second query.

We chose not to include an extensive content analysis of the web pages matched by the Google API. Instead, where patterns emerged further investigations were made through browsing the relevant web sites. This approach focuses on individual cases of significance,

rather than gaining a global picture, and has been described as 'significant anomaly identification' (Thelwall & Price, forthcoming).

3.5.3 Results

3.5.3.1 Size of web sites

Table 3-10 gives the estimated number of pages that have been indexed by the Google API database. Whilst there is a wide variety in the size of the web sites, there is some overlap; although the average size of an industry web site is smaller than both the average government and university web site, the largest industry web sites are bigger than the smallest government and university web sites.

Table 3-10 Estimated web site sizes as reported by the Google API on 30.06.2005

	Minimum	Maximum	Mean	Median
Government (39)	105	57,500	4,707	1,530
Industry (14)	1	3,756	870	489
University (13)	699	186,000	36,197	11,780

3.5.3.2 URL citation practices

Using the estimated number of hits from the Google API, Table 3-11 shows the mean number of URL citations from one organisational sector to another. There are wide variations with the strongest connections being amongst the universities and the weakest connections all involving the industry web sites. To show that the differences are not caused by large URL citation counts between a single pair of institutions, Table 3-12 shows the percentage of possible relationships that are reflected by at least one URL citation.

Table 3-11 The mean number of URL citations between organisations, by sector

		To		
		Government	**Industry**	**University**
From	**Government**	1.371	0.005	0.450
	Industry	0	0	0
	University	0.673	0.126	10.090

Comparing the interlinking of web sites at different levels, i.e., at the page level and the site level, has been used in other studies to find which provides the best indicator of research productivity (Thelwall, 2002e). Table 3-12 reinforces the pattern exhibited in Table 3-11.

Table 3-12 The percentage of possible relationships utilised

		To		
		Government	**Industry**	**University**
From	**Government**	215/1482=14.5%	3/546=0.5%	52/507=10.3%
	Industry	0	0	0
	University	79/507=15.6%	11/182=6.0%	102/156=65.4%

The variation in the average number of URL citations between different organisations cannot be attributed solely to excessively large URL citation practices between any particular pair of institutions.

3.5.3.3 Government-to-government URL citation practices

There was a relatively high mean number of URL citations between the thirty-nine principal local authorities, although these figures have been skewed by certain councils that were found to URL cite more highly than others. Of the 1,482 relationships that could be expressed via a URL citation from one institution to another (each of the 39 local authorities could have a relationship with each of the other 38 local authorities), 1,267 did not have any URL citations.

Figure 3-4 shows the interconnections between the local authorities' web sites, with the size of the arrows being used to indicate the number of URL citations represented. The organisations' positions on the diagram have been determined by the Kamada and Kawai (1989) algorithm, which tries to position connected sites close together although ignoring the strength of interconnection between a pair of organisations (Leydesdorff & Vaughan, 2006). Whilst *birmingham.gov.uk* can be seen to be the centre of the network, with URL citations to 29 of the 38 other government web sites; 1,693 of the 2,032 inter-government URL citations are to or from *warwickshire.gov.uk*. The large number of web URL citations from *warwickshire.gov.uk* to the district councils who cover the same area is due in part to an extensive database of links detailing all council services in the district, including those provided by the district councils. These links are often for information purposes, rather than an expression of the two organisations working together.

Table 3-13 County councils with their respective district councils

County Council	District Councils
.worcestershire.gov.uk	.bromsgrove.gov.uk; .malvernhills.gov.uk; .redditchbc.gov.uk; .cityofworcester.gov.uk; .wychavon.gov.uk; .wyreforestdc.gov.uk
.warwickshire.gov.uk	.warwickdc.gov.uk; .northwarks.gov.uk; .nuneatonandbedworth.gov.uk; .rugby.gov.uk; .stratford.gov.uk
.staffordshire.gov.uk	.cannockchasedc.gov.uk; .eaststaffsbc.gov.uk; .lichfield.gov.uk; .lichfielddc.gov.uk; .newcastle-staffs.gov.uk; .sstaffs.gov.uk; .staffordbc.gov.uk; .staffsmoorlands.gov.uk; .tamworth.gov.uk
.shropshire.gov.uk	.southshropshire.gov.uk; .bridgnorth.gov.uk; .northshropshiredc.gov.uk; .oswestrybc.gov.uk; .shrewsbury.gov.uk

The clustering of the web sites in Figure 3-4 corresponds with the two-tier structure of some local authorities (see Table 3-13). Whilst some local authorities are unitary authorities, others work on a two-tier system with the county council providing certain services, and district councils providing other services. The web sites of the county councils can be seen as being more connected with the other county councils and unitary authorities than the district councils that cite predominantly the URLs of other district councils under the same county council.

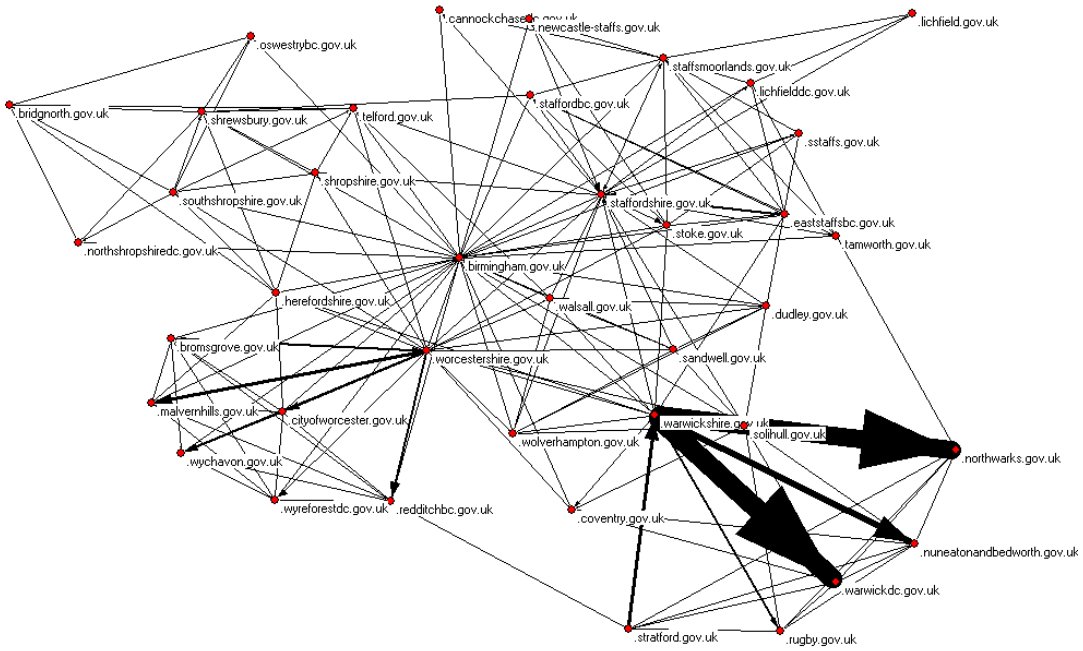

Figure 3-4 URL citations between local authorities' web sites in the West Midlands

3.5.3.4 Government-to-industry URL citation practices

Amongst the web sites researched there were only three URL citations found from a government web site to an industry web site: one from *warwickshire.gov.uk* to *jaguar.com*; one from *birmingham.gov.uk* to *mira.co.uk*; and one from *malvernhills.gov.uk* to *morgan-motor.co.uk.*

Inspection of the relevant web pages found that the hit for *warwickshire.gov.uk* was an email address in a database of local businesses rather than a link; bizarrely this was the email address of a chiropodist, and it was unclear why the email address was in the jaguar.com domain. The link to *mira.co.uk*, was from a library information services database of local businesses, and did not express any particular relationship. The *malvernhills.gov.uk* link to *morgan-motor.co.uk* did reflect the type of relationship that could be expected between local government and business; it was on a section of the site with the aim to promote the region, and promote partnerships between business and the community.

3.5.3.5 Government-to-university URL citation practices

Whilst there are a number of URL citations from local authority web sites to university web sites, there were less than for government-to-government URL citations. Nineteen local authorities didn't cite a single university, and no local authority cited the Newman Higher Education college web site, *newman.ac.uk.* Whilst it is not surprising that it is not cited as much as the large, prestigious universities, it is surprising that no URL citations were found from the local government web sites, especially as it is on the outskirts of Birmingham, and the institutions in the metropolitan borough of the West Midlands form the centre of the graph. Of the 507 relationships that could be expressed with web URL citations only 52 were.

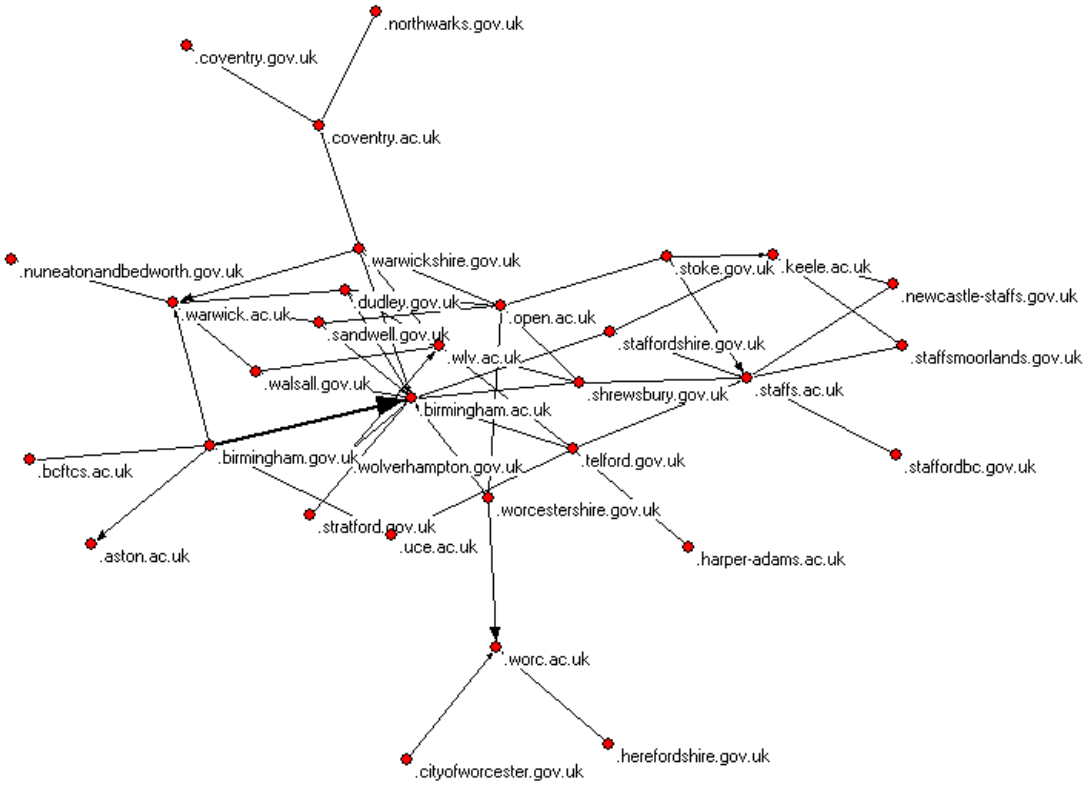

Figure 3-5 Government-to-university URL citations

As can be seen in Table 3-5 once again the URL citations follow geographic trends.

3.5.3.6 Industry URL citation practices

No URL citations were found from any of the organisations in the industry sector to any of the other organisations in the study.

3.5.3.7 University-to-government URL citation practices

The results show higher education institutions to be the most outward looking of the three types of organisation. Only the Birmingham College of Food, Tourism and Creative Studies, *bcftcs.ac.uk*, didn't cite any local authorities. Twelve local authorities' web sites weren't cited by any West Midlands university: *bridgnorth.gov.uk, northshropshiredc.gov.uk, oswestrybc.gov.uk, shropshire.gov.uk, southshropshire.gov.uk, lichfielddc.gov.uk, northwarks.gov.uk, nuneatonandbedworth.gov.uk, rugby.gov.uk, malvernhills.gov.uk, redditchbc.gov.uk, wychavon.gov.uk.*

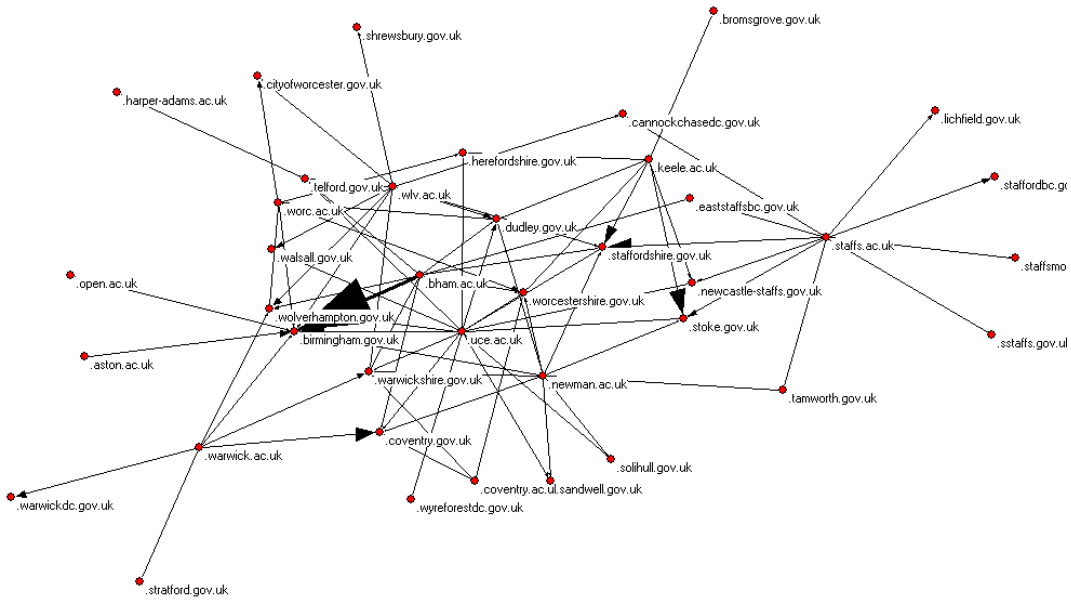

Figure 3-6 University-to-government URL citations

3.5.3.8 University-to-industry URL citation practices

There was a small amount of URL citing from university web sites to industry web sites. Seven of the universities didn't cite any of the industry sites and eight of the industry sites weren't cited. Whilst a variety in citing practices could be expected due to different universities having differing academic focuses; there are also universities with specific automotive courses, but no URL citations. For instance Aston University has a high research rated engineering department, with a course in automotive product design but the Google API found no URL citations from *aston.ac.uk* to any of the industry web sites in this study. This affirms that URL citations do not fully reflect real-world connections.

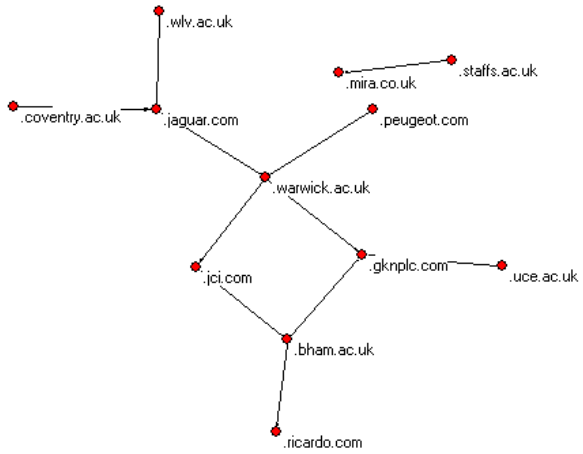

Figure 3-7 University-to-industry URL citations

3.5.3.9 University-to-university URL citation practices

Unsurprisingly the university sector is highly interconnected, although two universities don't cite any of the other universities: Birmingham College of Food, Tourism and Creative Studies, and Harper Adams University College. The mean number of institutions the other eleven universities cite is 9.27. All universities are cited by at least three other universities.

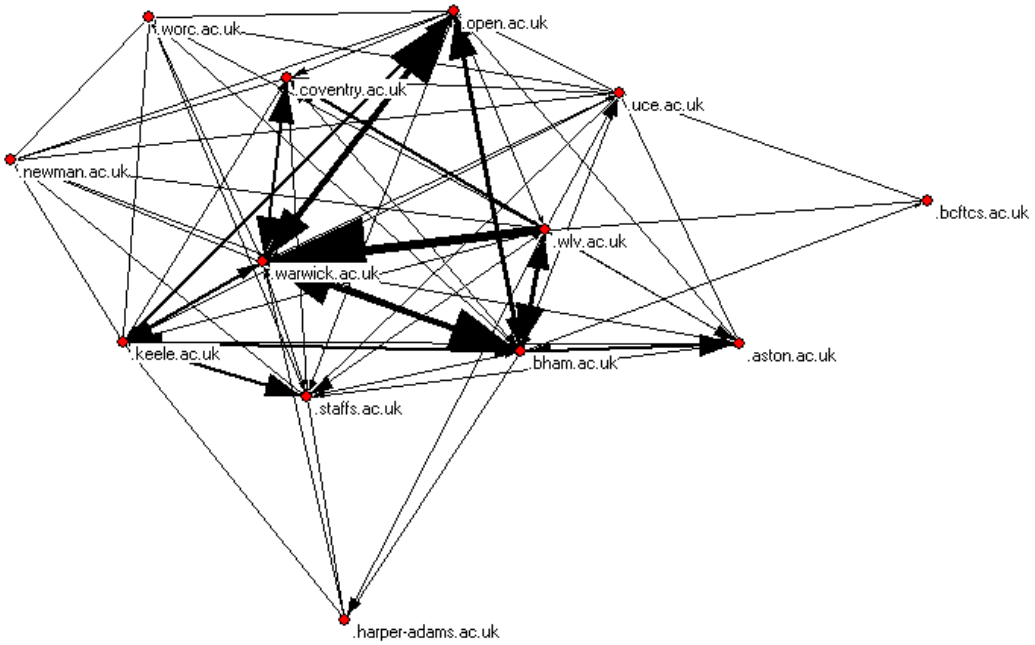

Figure 3-8 University-to-university URL citations

3.5.4 Discussion

The results show that there are major differences in the URL citation practices of web sites in the university, industry, and government sectors, as would be expected between heterogeneous organisations (Vaughan & Wu, 2004). Most notable is the lack of URL citations both to and from the industry web sites that have been found using Google API. It is not surprising that there are fewer URL citations from the industry web sites, especially to competitors, as they have much to gain by increasing the stickiness of their web site rather than encouraging surfers to move on to other web sites (Shaw, 2001). That no URL citations were found from any industry web site, to any of the local government or university web sites shows that URL citations are not a suitable way of indicating collaboration, especially in the context of an EU funded, university-based project that gave web sites to automotive supply chain companies in the area (Costello, Garner, Homer, & Thompson, 1999). The weak network reinforces the findings of Heimeriks et al. (2003). Their study did not address the differences in linking behaviour of the different sectors, although it was noted that the linked network was an unrelated set of organisations. This is attributable to the proportion of the web sites that were crawled in their study producing even fewer results, which may have been exacerbated by the differences that may arise due to URL citations being searched for rather than web links.

Whilst the lack of URL citations could be due to an active attempt to keep web users on the web site, it could also be caused indirectly due to the type of information placed on the web, and who places the information. It could be that the collaboration that occurs between an organisation and local government and universities are not seen as worthy as highlighting on a web site that will be seen internationally, or that the collaborations are mentioned, but do not take the form of web URL citations.

It is also possible that the URL citations exist, but there are restrictions in the indexing of the pages: organisations may not want their pages indexed, and use the robots.txt protocol; Google may choose not to index the pages; or the pages may not be in a format that can be read and indexed by the Google robot. Whilst it is the case that the industry web sites have fewer pages indexed by Google API, there are still a significant number of pages indexed. It seems unlikely that technical reasons such as these could account for the large differences in URL citation practices found. It is more likely that web sites for the three sectors have different priorities. Table 3-14 gives a possible simple type variation.

It is also possible that there were a number of URL citations that were not found as they were to domains that were not incorporated within the study. Organisations in the industry sector may have a diverse web presence, with many different web sites for different subsidiaries or for different purposes. This raises the question of what comprises an organisation and which web sites should be investigated. Affiliations between web sites are often established to emphasise the credibility of the web sites, with financial web sites occupying a central position in the network (Park, Barnett & Nam, 2002).

Whilst more URL citations were found from local government and universities to the industry web sites, these were still relatively few. This is worrying and highlights quite a disconnected web. That seven of the fourteen industry sites weren't URL cited by any of the local authorities or universities is not particularly healthy from the organisations' perspectives. Even if a lack of web URL citations does not reflect a lack of collaboration, it does reflect a significant lack of web presence, something that is important to the promotion of their own organisations, and the region.

Table 3-14 Suggested web site types for university, industry and government sectors

Sector	Web site main functions	Possible common uses for links and URL citations
University	Communication between individuals and groups; providing access to useful online information; promoting the organisation (Middleton et al., 1999).	Pointing to other information sources; acknowledging collaboration.
Industry	A marketing tool (Shaw, 2001) providing commercial and non-commercial information, entertainment; aiding transactions (Huizingh, 2000).	Accessing services such as online payment forms; reflecting organisational hierarchies.
Government	Promotional; content provider; aiding transactions (Musgrave, 2004).	Pointing to information of use to the public; reflecting organisational hierarchies.

Local governments' concern with local issues is highlighted by their URL citation patterns. The directional graph showing the citing between the different local authorities shows clusters of the district councils, with their respective county councils, as well as citing between those authorities that are next to each other. URL citation patterns from government to universities also follow a geographic pattern. Analysis of the URL citations finds that they do not necessarily reflect relationships between the organisations, but are often used as sources of information and the numbers are very much reliant on the URL citation or linking policy of the individual site. It is likely that the URL citation patterns would differ for government institutions at a regional level.

The university sector was by far the most heavily connected; more outward looking to the other sectors, and heavily connected to the other institutions within the university sector. This is not unexpected, both because of the early adoption of the internet by the academic sector, and their being involved in the dissemination of information. They are still relatively disconnected however. Note that the link model developed for university web site interlinking (Thelwall, 2002d) is symmetrical and so the asymmetric nature of links between sectors is a significant indicator that their web sites are used for different purposes.

As the numbers of web URL citations between the organisations is small, there is an increased need for transparency in the way search engines work, and how much or how often each organisations web site is indexed; just a handful of URL citations could change an organisation's position relative to others in the field. The limited number of URL citations between sites does mean that a random sample of pages could be selected that cite from one organisation to another. In the vast majority of cases there are less than ten URL citations between any two web sites, and where there are more than ten, it is less than the thousand the Google API allows to be retrieved.

When looking at a small selection of web sites, even within a limited geographical area, there are fewer connections than may be expected in the real world. Therefore it is necessary to look at larger numbers of web sites for a reliable picture to emerge, and with the number of queries necessary growing exponentially, it is not practical to use a similar methodology on a larger scale.

Previous studies have found that the majority of university outlinks do not show a working relationship between the organisations concerned, but rather are placed for other reasons such as pointing to information resources. Analysis of a small selection of the web URL citations from local authority web sites would seem to confirm the same type of relationships. That the URL citations still produce clusters of known real world relationships suggests that these relationships are reflected as URL citations, but not necessarily ones that explicitly state the relationship.

3.5.5 Conclusions

This exploratory study has investigated the potential use of web URL citations as weak benchmarking indicators to determine levels of collaboration between organisations within the different sectors of the Triple Helix model. The results show that URL citation only partially reflects real-world relationships because of different purposes for web sites between sectors, and also because there is no imperative to create links or URL citations to advertise a collaboration. It seems likely that the use of web links and URL citations would be as one variable used in conjunction with one or more other manifestations of relationships between organisations, and that web-based indicators will always give only a partial picture of collaboration, particularly collaboration concerning industry.

There is a need for more research into the web use of the non-academic community: how they utilise their web sites and their reasons for link placement. The differences between the university, industry, and government sectors have been emphasised, and a far more extensive classification of web sites is likely to be needed if meaningful assertions are to be made about what sort of weighting can be ascribed to different web links or URL citations. The weighting of web links is clearly needed, as a link from some sites is clearly of more value than a link from another.

Even if enough information was known for appropriate weightings to be ascribed to the different meanings of web links or URL citations, it would be necessary for other manifestations to be taken into account. Whilst URL citations provide useful information, they are not enough on their own. The small number of connections between the web sites means that it is necessary for subtler connections to be analysed, and future research into the triple helix manifestations on the web may find text analysis rather than link analysis or URL citation analysis a more productive avenue to follow: searching for occurrences of one organisations name on another organisations web site; or co-occurrences of organisations names on other web sites.

3.6 University-industry-government relationships manifested through MSN reciprocal links

3.6.1 Introduction

Webometric investigations are restricted by the data collection tools available. As data collection tools improve it is necessary to re-examine previous investigations to determine whether alternative methodologies are available. The new Live Search (formerly MSN) search engine operator, *linkfromdomain:*, introduced in October, 2006 (MSDN, 2006), enables the manipulation of web data in a way that was previously only possible with a web crawler. Before, if information was required regarding a site's outlinks as well as inlinks, it was necessary to crawl a group of web sites. Without the crawling of web sites the outlink

information available was restricted to the searching for specific URL citations or second/top-level domain names, e.g., *"co.uk" site:wlv.ac.uk.* As such, it was impossible to get a list of outlinks to those sites that shared a second/top-level domain name with the site under investigation, as it would list all the pages of the web site that included a self-link. Whilst self-links can be excluded when measuring inlinks, by excluding all the site's own pages with the –*site:* operator, the mutually exclusive argument would always produce zero results.

As well as inlinks, and outlinks, the new command also allows the investigation of reciprocal-links: where the first web site is found to link to a second web site, and the second web site is found to link back to the first web site. Again, previously such investigations were restricted to a small group of web sites that could be crawled by the investigator with a web crawler, or the links between specific web sites could be investigated with a search engine.

From the perspective of investigating the relationships between organisations through their web manifestations, the ability to identify relationships with organisations that are unexpected, and the ability to identify reciprocal-links, seems to have obvious advantages. It seems reasonable to presume that if two organisations are both linking to one another it is less likely that the links have been placed for arbitrary reasons, than if only one of the organisations is linking to another.

This study investigates what MSN reciprocal links can tell us about the relationships between organisations by re-examining the case of the UK's West Midlands automobile industry, the focus of the previous investigation. The original study was carried out before the new operator was made available.

This investigation (a) changes the data type of the previous study from URL citations to reciprocal-links and (b) extends the scope of the connections to include sites outside of the core set investigated, using the following questions.

- What kind of university-industry-government collaborations, if any, are reflected by MSN reciprocal-links?
- Do MSN reciprocal-links identify more collaborative relationships than previous methodologies?

3.6.2 Research methodology

This study investigates the relationships between the UK's West Midlands automobile industry as exhibited by MSN reciprocal-links. The study has two main parts:

1. Data collection.
2. Classification of a selection of reciprocal-links.

3.6.2.1 Data collection

As with the previous investigation the web sites used for this investigation are the 13 higher education institutes in the West Midlands region (West Midlands Higher Education Association, 2006), the 14 web sites identified by the automotive industry unit of the Department of Trade and Industry as having a significant role in the region (DTI, 2006), and the 38 West Midlands local councils identified on the government web site (DirectGov, 2006).

A program was written to determine those web sites that each of the 'seed' URLs had reciprocal-links with. For each of the seed URLs a query was sent to the MSN API as many times as necessary to retrieve either all the reciprocal web sites identified (if less than 1,000),

or the MSN imposed maximum of 1,000 where more than 1,000 reciprocal web sites were identified.

e.g., linkdomain:wlv.ac.uk linkfromdomain:wlv.ac.uk –site:wlv.ac.uk

Where more than one URL was identified for an organisation both URLs were used.

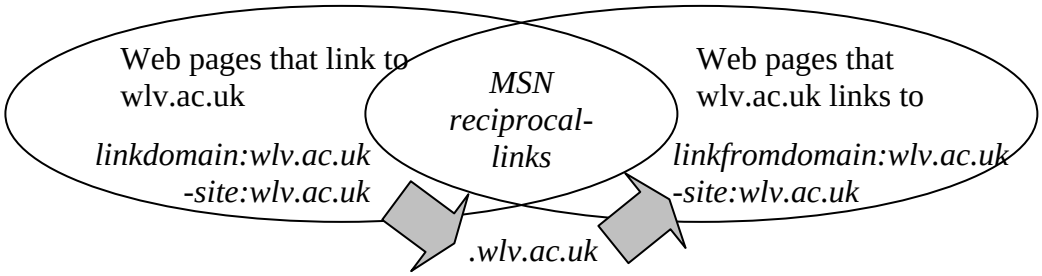

Figure 3-9 MSN reciprocal-links

Since it is the first use of the *linkfromdomain:* operator it is necessary to briefly discuss the results that are provided when the operator is used on its own and in conjunction with the *linkdomain:* operator. The *linkfromdomain:* operator provides a list of pages that are linked to from the domain name entered. Whilst it has been reported that it doesn't provide a list of outlinks for sub-domains (SearchEngineWatch, 2006), further investigations indicated that whilst it does not currently allow a search to be restricted to sub-domains, those pages within the sub-domains are indexed and included in the search at the domain name level, e.g., *linkfromdomain:wlv.ac.uk* will return those pages linked to by pages in the sub-domain *linkanalysis.wlv.ac.*uk even though the query *linkfromdomain:linkanalysis.wlv.ac.uk* would produce no results.

Combining the *linkfromdomain:* operator with the *linkdomain:* operator for the same web site will produce a list of pages which are both linked to from the domain, and link to the domain. It should be noted however that MSN reciprocal-links are not necessarily reciprocated; there is still a perspective that needs to be taken into consideration. For example, one of the hits for the query "*linkdomain:wlv.ac.uk linkfromdomain:wlv.ac.uk"* is the web site *innovation-direct.com*, but when entering the query "*linkdomain:innovation-direct.com linkfromdomain:innovation-direct.com"* the *wlv.ac.uk* web site is not found.

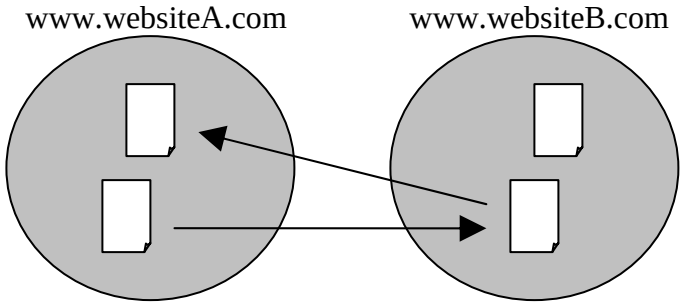

Figure 3-10 Example of non-reciprocated reciprocal-links

Figure 3.10 provides an example of non-reciprocated reciprocal-links. Whilst the query "*linkdomain:websiteA.com linkfromdomain:websiteA.com"* would return one of web site B's pages amongst the hits, "*linkdomain:websiteB.com linkfromdomain:websiteB.com"* would not

return either of web site A's pages, as there are no pages on web site A that are both linked to by web site B and link to web site B. In the alternative document model terminology (Thelwall, 2002d), web site A is aggregated at the site document level whereas web site B is aggregated at the page level.

A program was also written to determine the number of different web sites that had MSN reciprocal-links from the original seed web sites. The notion of the 'web site' may be operationalized in a number of ways: lexically, semantically, or topologically (Cothey et al., 2006). For the purposes of this investigation the web site is operationalized on a lexical basis. The retrieved URLs were 'cleaned' by removing anything following the top-level domain name and then deleting duplicates. Due to the large number of web sites with shared web hosting arrangements reflected in their domain names, web sites with different sub-domain names were treated as different web sites, e.g., barcelonaforbeginners.blogspot.com was considered to be a separate web site to webometrics.blogspot.com. Calculating the number of different organisational web sites without looking at each of them individually can only produce a rough estimate: in this case sometimes separating sub-domains that are part of a common larger web site, and sometimes combining web sites that should be considered separate due to a different hierarchical structure. For example, http://www.myspace.com/warwickuniversity and http://www.myspace.com/caseyleaver would be combined even though most people who visited them would surely agree they are separate web sites.

3.6.3 Classification of relationships between the organisations

The relationships between 50 MSN reciprocally linked web sites were investigated for each of the sectors: academic, industry, and government (where there were 50 reciprocally linked organisations which were identified). To prevent bias towards the larger organisations the 50 reciprocally linked organisations were taken from a subset based on up to ten reciprocal-links for each of the organisations in the study.

The classification scheme addresses the question of whether a collaborative relationship can be determined between the seed organisation and the organisation or individual whose web site it has a reciprocal-link with: whilst it is not a classification of the individual links, the individual link and linked to page were the starting points of the content analysis. The broad categories of 'collaborative relationship' and 'non-collaborative relationship' include the types of relationships identified under those headings in the earlier study into academic web links and collaboration, a collaborative relationship includes: research collaboration; one party can carry out work on behalf of another; one party providing sponsorship or funding; the two organisations working together towards mutual goals; as well as informal relationships, such as speaking at a conference. A non-collaborative relationship includes: use of information resources and references; acknowledgements; research support; tourist information.

The more detailed break down of 'reflecting a collaborative relationship' reflects the policy of many organisations utilising different web sites for different groups and services. A reciprocally-linked web site may be part of the same organisation (i.e., a subsidiary web site), an external organisation (i.e., a non-subsidiary web site), or a web site that is the result of a collaboration with another organisation. Thus a subsidiary web site may actually be owned by several collaborating organisations. Whilst the boundaries of an organisation are not always

clear, in this study the boundary is determined by whether the web site contributes to the official work of the organisation.

Table 3-15 Classification scheme of relationship between organisations

Type of relationship
1. Collaborative relationship (formal or informal)
a. Subsidiary web site
b. Subsidiary web site – with partners
c. Non-subsidiary web site
2. Non-collaborative relationship

3.6.4 Results

As can be seen from Table 3-16 a wide variation was found amongst the number of sites that each of the seed web sites had reciprocal-links with.

Table 3-16 Number of reciprocally linked web sites

	Minimum	**Maximum**	**Mean**	**Median**
University	2	508	124	41
Industry	0	10	2	1
Government	0	67	11	8

The estimated number of MSN reciprocal-links is not a static number, but rather varies according to how deep the results are retrieved from, apparently becoming more accurate the deeper the results are retrieved from. For example, when asking the API to return the first 50 hits that fulfil the query: "*linkdomain:warwick.ac.uk linkfromdomain:warwick.ac.uk –site:warwick.ac.uk"* it estimates the total number of results to be 3,251. When the same query is entered requesting the results from 351-400, the estimated number of results has fallen to 516. The actual number of results that could be retrieved from the database was 508. The results shown in Table 3-16 are based on the actual number of URLs which were extracted.

1,607 different web sites were identified as having MSN reciprocated links from the initial seed web sites.

Table 3-17 Types of relationship between MSN reciprocally linked organisation

	Type of 'Seed' Organisation		
	University	**Industry**	**Government**
1. Collaborative Relationship			
a. subsidiary web site	13/50	23/30	18/50
b. subsidiary web site – with partners	2/50	1/30	2/50
c. non-subsidiary web site	22/50	4/30	24/50
2. Non-collaborative relationship	13/50	2/30	6/50
Percentage of collaborative relationships	**74%**	**93%**	**88%**

The low number of MSN reciprocal-links retrieved for the organisations in the industry sector meant that it was not possible to classify fifty relationships; instead all thirty identified relationships were classified.

3.6.5 Discussion

Previous investigations into the ability of web links to provide indicators of collaboration between organisations suggested the low proportion of links reflecting collaboration would necessitate too much human intervention to make it a viable course of study. However, this pilot investigation of MSN reciprocal-links, enabled through the recent introduction of the *linkfromdomain:* operator, suggests that further investigation is necessary before the notion of links as indicators of collaboration is cast aside.

3.6.5.1 What kind of university-industry-government collaborations, if any, are reflected by MSN reciprocal-links?

As would be expected, this investigation found MSN reciprocal-links to a large number of web sites outside the original seed web sites, even though MSN reciprocal-links may be considered a fairly restrictive type of relationship in comparison to the existence of a single inlink or outlink. This, coupled with the high proportion of the MSN reciprocal-links that were found to reflect collaboration between organisations, far higher than in previous investigations of outlinks, suggests that there is a role for web links in the investigation of relationships between organisations.

Whilst web sites from all three sectors have MSN collaborative links, there are clear differences in both the type of collaborative relationship that is reflected, and the percentage of MSN reciprocal-links that reflect collaboration. The most noticeable difference in the type of collaboration that is reflected by MSN reciprocal-links is the high proportion of MSN reciprocal-links in the industry sector that are to subsidiary web sites. The lack of outlinks in the industry sector to external organisations has been noted in previous studies (Shaw, 2001), and it is found again in this study. It seems likely that the different linking practices of the different sectors needs to be reflected in different data collection methodologies. Whereas an analysis of outlinks for university and government web sites is likely to produce too many links to be of any use when investigating collaboration, it may be a more appropriate data collection technique for industrial organisations.

This study also highlights the large number of subsidiary web sites. Part of the reason for this is the large number of individuals that were using the facilities offered by blogging web sites, and other web publishing sites, for official organisational work. Such use is unsurprising: little skill is necessary in setting up such web sites, and it bypasses the official bureaucracy that may impede publication on the primary organisational web site. As well as having implications for the calculation of impact factors for web sites, they also cause dead-ends when investigating collaborative networks as they generally use a sub-domain of a blog site, and as has already been mentioned, the *linkfromdomain:* operator does not currently work for sub-domains.

It seems likely that the differences in the proportion of MSN reciprocal-links that reflect collaboration is a sign of the size and scope of the web sites based in the different sectors. The idea that reciprocal-links are more likely to connect collaborative organisations than a single outlink, is based on the supposition that when sites are linking to one another it is

less likely to be for non-collaborative reasons. The larger the web site and the more diverse the content held on it, the more likely the returning link is for a non-collaborative reason.

A potential weakness in using reciprocal-links to find collaborations between different organisations is the large number of intra-organisational links which are showing up as reciprocal-links between different web sites; how can intra-organisational relationships be distinguished from inter-organisational relationships? Whilst it seems likely that satellite web sites have a different linking structure with their main organisational web site than with other organisational web sites, this is an area which needs further investigation.

3.6.5.2 Precision, recall and bias

When comparing different methods, one of the standard measures is recall: the percentage of correct results that have been retrieved out of the total expected answers (Mikheev, 2003). As the total expected answers are unknown for documents on the web, for this paper, the recall question becomes: will MSN reciprocal-links identify more collaborative relationships than previous methods? As MSN reciprocal-links are a subset of a web site's outlinks, the answer to the question is no; certain collaborations will be reflected by outlinks that aren't reflected by MSN reciprocal-links, whilst in theory there will be no collaborations that are reflected by reciprocal-links that aren't reflected by outlinks (in practice the volatility of search engines means that a site may be returned for a reciprocal-link that wasn't returned for an outlink). However, the quantity of information that can be gathered by combining the *linkfromdomain:* and *linkdomain:* operators in the MSN API, and the high proportion of MSN reciprocal-links that reflect collaboration means that a large number of collaborations between organisations may be identified quickly and simply; more than would have been possible with previous methodologies.

It is generally the case in information retrieval that an attempt to increase levels of precision (i.e., the percentage of correct results in produced results (Mikheev, 2003)) comes at the expense of recall, and whilst MSN reciprocal-links decrease recall, the results above show a relatively high level of precision (at least in terms of reflecting collaborative relationships between web sites if not between organisations). In addition, it is important to consider the issue of bias: those types of relationship that are unlikely to produce MSN reciprocal-links. If the objective of a research project is to identify collaborations and the method used is MSN reciprocal-links then what type of sampling bias would be introduced?

Whilst the web is used extensively by all sectors of society, it is not ubiquitous. Many organisations still do not have web sites, and amongst those that do there are vast differences in the regularity with which they are updated, the amount of information they provide, and who has permission to change the information provided. It seems likely that collaboration in the form of MSN reciprocal-links is going to be more visible where there is a strong focus on electronic collaborations and communications.

3.6.6 Conclusion

The purpose of this pilot study was to establish the feasibility of utilising MSN reciprocal-links to investigate collaboration between organisations. It has shown that whilst there are limitations in the use of search engines for webometric investigations, and specifically limitations with MSN's *linkfromdomain:* operator, a lot of potentially useful information is now available through MSN's search engine which can tell us a lot about the collaboration between different organisations.

There is still a lot more research that is necessary before MSN reciprocal-links can be used as indicators of collaboration between organisations:

- How much of the collaboration that is within the web link structure is being lost by focusing on reciprocal-links rather than outlinks?
- How can intra-organisational collaboration be distinguished from inter-organisational collaboration?
- How great a hindrance is the operator's current inability to deal with sub-domains to mapping the collaboration network beyond the first generation of URLs created by the seed URLs?
- The feasibility of measuring the strengths of collaborations identified by reciprocal-links.
- Alternative methods of data collection for certain commercial organisations.

Hopefully, however, as more search engines increase the accessibility of their databases, webometrics will be able to provide more indicators.

4 Principal research design and methodology

4.1 Introduction

The aim of this thesis is to determine whether link analysis of the web can provide a new source of information about knowledge-based innovation systems in the UK, assuming that interactions between organisations from different sectors enables greater innovation by the system as a whole (Potratz & Widmaier, 1996). The web cannot, however, be expected to provide a single indicator of an institution's 'triple-helix-ness' in the way Boudourides et al. (1999) envisaged. The set of organisations that comprise a knowledge-based innovation system are necessarily heterogeneous: the web is used differently by different organisations within different fields (Kling & McKim, 2000) and within different sectors of society (Middleton et al., 1999; Shaw, 2001; Musgrave, 2004). Such indicators are liable to reflect the way types of organisation use the web rather than their off-line interactions. As Vaughan and Wu (2004) stated, correlations are not likely to be found for heterogeneous groups of organisations; in the same way as patents should only compare the performance across a single technology (Narin, 1993). Comparison of organisations from different sectors also lacks a universal indicator against which to compare, an important aspect in validating any conclusions (Thelwall, 2004a).

Despite the inability of the web to provide comparative indicators of organisational innovativeness, it can nonetheless provide information about the collaborations between organisations, collaborations that have been described as so widespread within innovative companies as to seem essential to the innovation process (Smith, 2005). The extensive classifications of the web links placed on university web pages shows that a significant proportion, if not a majority, reflect a collaboration between the linked organisations (see section 3.3 & 3.4), and when investigating the interlinking between a related set of organisational web sites clusters emerge of known real world relationships (see section 3.5).

Working within the recognised limitations of link analysis to provide a new source of information about knowledge-based innovation systems within the UK, the final study investigated the link relationships of the UK's pharmaceutical industry. This chapter starts with the proposal of a number of hypotheses about the ability of the web to provide new information about knowledge-based innovation systems based on both the findings of the literature review and the preliminary investigations. This is followed by a detailed description of the methods adopted within the final study to test the hypotheses.

4.2 Hypotheses

4.2.1 A search engine API can be suitable for data collection

Data collection has been a significant issue within the preliminary investigations of this thesis, as well as within many previous webometric studies. The solution is inevitably one of compromise as it is impossible to collect a true picture of the web for anything but the most limited of studies. It is therefore necessary to find a method that sufficiently meets the criteria of the investigation without causing excessive disruption to the web sites under investigation.

The most appropriate data collection tool for collecting information about the interactions between web sites that form knowledge-based innovation systems seems to be a search engine with an API. This is principally due to the need to include a large number of

organisations. The necessity of including a large number of web sites was shown in the initial pilot investigation into URL citations between the automobile industry, universities, and local government in the West Midlands (see sections 3.5), where a weaker than expected network of linking between the organisations was found. Whilst it was expected that the industry sector would have less outlinking than other sectors (Shaw, 2001), the total lack of outlinks to the other web sites in the investigation demonstrates the need for the inclusion of a larger set of web sites, some of which may be unknown at the beginning of the investigation. A large number of web sites, however, excludes the use of a web crawler if the investigation is to be carried out in an ethical manner (see section 3.2). This leaves a search engine as the only feasible option. Although there have been objections raised about the suitability of search engines in webometric investigations (Snyder & Rosenbaum, 1999), it is thought that the UK web space is sufficiently crawled by the major search engines to provide a useful source of information (Thelwall, 2001d), and the operators and accessibility provided by the search engines, in particular Live Search gives sufficient access to this information for it to be a useful source of information.

- **H1 -** The UK web space is sufficiently well crawled by Live Search for use in webometric investigations into collaboration.
- **H2 -** The operators and accessibility currently offered by Live Search can provide useful information about the web links between web sites.

4.2.2 Classification is necessary for the identification of collaborative web links

Classification studies have shown web links to be placed for a wide range of reasons, from navigational and social, to purely gratuitous (Thelwall, 2003d). Investigating the potential of web links as an indicator of collaboration between the two actors represented by the two linked pages in the preliminary investigations have echoed the findings of previous investigations (i.e., Vasileiadou & van den Besselaar, 2006): whilst there is a significant enough proportion of web links placed for collaborative reasons (see section 3.3 & 3.4) to make them an area worthy of investigation about organisational collaboration, most are placed for information purposes. However, the proportion of links that have been found to represent a collaborative relationship was much higher when investigating MSN reciprocally linked organisations, particularly for organisations within the industry sector (see section 3.6).

- **H3 -** A significantly high enough proportion of outlinks and MSN reciprocal-links reflect collaboration for organisations within the industry sector for them to provide a useful tool in investigating an organisation's collaboration.
- **H4 -**The majority of links placed on the web are placed for highlighting information rather than reflecting collaboration. As such the most central organisations within a web network will be those that play a role in the dissemination of information rather than those that are highly collaborative.

4.2.3 Web data about collaboration is different from traditional sources of organisational collaboration

The web is likely to prove to be a new source of information, rather than a repetition of the information available within traditional bibliometric databases. The web provides the opportunity to show informal relationships (Wilkinson, Harries et al., 2003), collaborations

that are still in progress (Bossy, 1995), and those that are not necessarily novel (Meyer & Bhattacharya, 2004).

The web is unlikely, however, to include all the information available in traditional databases. Whilst it is a place for both formal and informal collaborations, such collaborations are becoming increasingly blurred (Barjak, 2006). Previous attempts have failed to find co-authorship reflected in links between personal homepages (Kretschmer & Aguillo, 2004; Kretschmer et al., 2007), and the dynamic real-time nature (Ingwersen, 1998) means that collaborations that were once visible on the web may disappear, whilst they may continue to appear within bibliographic databases.

Whilst Heimeriks et al. (2003) mapped both communication and collaboration networks, their investigation looked at different characteristics of the communication network; whereas their investigation compared the co-inlinks and bibliographic coupling, and investigated the network as a whole, this investigation compares those web links that are found to reflect a collaborative relationship with the appearance of multiple organisational names within scientific papers and multiple assignee names within applied for patents.

- **H5 -** Web links can reflect collaboration between two organisations not visible through traditional bibliometric sources.
- **H6 -** Traditional sources can provide information about the collaboration between two organisations not visible on the web.

4.3 Methodology

The final study in this investigation into the web manifestations of knowledge-based innovation systems in the UK tests the above hypotheses with a link analysis of the UK's pharmaceutical industry. This study is based on the network of web links derived from the organisational members of the Association of the British Pharmaceutical Industry. Using link data collected through the Live Search API: Social Network Analysis (SNA) methods are applied to the network of web sites to identify the central organisations; web links are classified to determine whether or not they reflect a collaborative relationship; and the classification results are compared with those obtainable through the traditional bibliometric sources of patents and science papers to investigate the extent that such information is new.

4.3.1 Population selected in this study

The final study investigates the collaborative relationships reflected in the web links of the members of the Association of the British Pharmaceutical Industry (ABPI), the pharmaceutical industry's main industry body and lobby group in the UK (CorporateWatch, 2003).

The pharmaceutical industry has been the focus of many previous bibliometric investigations, analysing both patents and scientific articles (e.g., Koenig, 1983a; 1983b; Vinkler, 1994; McMillan & Hamilton, 2000), and it is a field where the complex nature of the interactions between the different sectors of society was recognised long before the terms Mode 1 and Mode 2 science were coined (i.e., Koenig, 1983a). The pharmaceutical industry is a heavily research dependent field (Koenig, 1983b) with public science having an important role within the industry sector (McMillan & Hamilton, 2000). There is also a high level of patent enforcement that encourages patenting (McMillan, 2000) and the government is known to have an important regulatory role. It is also to be expected that pharmaceutical organisations' web sites will be more than a shop window for their products as they are often a

destination by people looking for information about diseases, diagnoses, and medications (Maddox, 1999), and to satisfy potential customers it is necessary that they provide that information. Trust is especially important in the provision of information about drugs, and promotion of collaborations with other organisations is one way that pharmaceutical organisations can build-up the public's trust in their products as affiliations between web sites are often established to emphasise their credibility (Park et al., 2002). It has been found that two of the features that users take into consideration when assessing the quality of medical information on a web site are the citation of scientific resources and whether the web sites are an official source (Eysenbach & Köhler, 2002). If web links are not found to be a useful source of information about collaboration within the pharmaceutical industry, it seems unlikely that it will provide useful information about organisational collaboration in other fields.

The ABPI web site's list of full member organisations, general affiliate members, and research affiliate members, provides an authoritative list of 150 organisations with an interest in the UK's pharmaceutical industry, and includes a URL for the majority of the members' web sites. Such a list may be considered a sufficiently related collection of web sites for there to be a certain amount of interlinking between them, as well as sharing links to and from other unidentified organisations. Although it is part of the nature of knowledge-based innovation systems that they consist of organisations from different fields and different sectors of society working together, merely selecting a random sample of web pages would be unlikely to find many connections (Thelwall, 2003b). With membership the criteria for inclusion on the list, it allows for the inclusion of organisations that recognise themselves as having a role in the pharmaceutical industry, even if it is not immediately apparent to external organisations. For example, the general affiliate members include organisations that would not be primarily associated with the pharmaceutical industry such as the software giant Microsoft and the consulting company Accenture. In addition to listing the organisational members, the ABPI web site also has links to a number of other web sites: associated organisations, patient organisations, and 'other links'. For the purposes of this investigation only the member organisations were selected as the inclusion of the other organisations was not necessarily due to a choice on the organisation's part, but rather the whim of the site creator, and there are no details about the ABPI's inclusion policy.

Where a URL was provided for an ABPI member's web site it was checked to see that it was both accessible and referred to the organisation named on the ABPI web site. The Google search engine was used to identify the web sites of organisations that either did not have a URL listed with their contact details on ABPI's web site. If selecting the URL redirected the investigator to another site then both URLs were then associated with the particular organisation. Of the 150 organisations, 145 had web sites that were identified by this method.

4.3.2 Link data collection

Each of the three major search engines provides an API that enables the user to extract the search engine data in an ethical manner. There are, however, vast differences in the operators that the major search engines enable and the number of queries they allow to be sent. Table 4-1 compares the operability of the three major search engines at the time of collecting the data, in March 2007.

- Link – Within Yahoo and Google, this command finds pages that link to a specific web page. Although Live Search uses the *link* command its definition is indistinguishable from their *linkdomain* command (Live Search, 2007).
- Linkdomain – Finds pages that link to a particular domain.
- Linkfromdomain – Finds pages that are linked to from a particular domain.
- Site – Restricts the search to pages that are within a particular web site or top-level domain.

Soon after collecting the data for this investigation Live Search discontinued the use of its *link* and *linkdomain* operators due to misuse (Live Search, 2007).

Table 4-1 shows the facilities offered by Live Search to compare favourably with those offered by the other major search engines. It enables the greatest variety of link data to be retrieved whilst also allowing the most results to be retrieved. Whilst theoretically both Yahoo and Live Search allow the same number of results to be retrieved, 10,000 x 50 for Live Search and 5,000 x 100 for Yahoo, in actuality by asking for results in smaller chunks Live Search allows for the gathering of more results than Yahoo. For example, if there are only 34 more results to collect, Live Search would fail to use 16 of its potential results; Yahoo would fail to use 66. The Live Search index, however, is not as large as Yahoo's or Google's. A search for the term *pharmaceutical* within the ccTLD *.uk* finds 2,370,000 pages on Yahoo, 1,320,000 on Google, and only 608,809 on Live Search. It is necessary to include a keyword in addition to the *site* operator as it is obligatory within the Yahoo search engine, and the Google results have been found to be more volatile unless used in conjunction with a keyword (see section 3.5.2.1). The reasons for the large differences between the three search engines are unclear, as is whether or not there would be any additional advantage in incorporating the additional pages that are included within the Yahoo index or those included within the Google index. It may be that the larger indexes are the result of deeper crawling of a handful of large web sites, or the inclusion of numerous web sites not recognised by Live Search. Alternatively it may be that the search engines have different levels of accuracy in their estimations.

Table 4-1 Comparison of the facilities offered by the three major search engines' APIs

	Google	Live Search	Yahoo
Number of Queries	1,000 per license per day (10,000 with permission)	10,000 queries per day per IP address	5,000 per day per IP address
link	✔	✘	✔
linkdomain	✘	✔	✔
linkfromdomain	✘	✔	✘
Site	✔	✔	✔
Records per query	10	50	100
Results that may be downloaded	1,000	1,000	1,000

Although the numbers of inlinks, outlinks, and reciprocal-links that a web site has are not being investigated within this study for determining indicators of an organisation's success or innovativeness, they are collected along with the actual link data so that the suitability of the data collection method can be discussed, in terms of: the proportion of the identified links that

are accessible, and the extent to which the web sites included in the study are crawled. It has previously been found, however, that there are inconsistencies in the results of search engines, most noticeable in the addition of two mutually exclusive searches which between them should add up to all the pages indexed within a particular web space (Smith, 1999a). Table 4-2 shows the discrepancies between the results provided by the search engines and the results expected to be found by the search engines. Although theoretically the addition of those web pages that include the term *pharmaceutical* that have the ccTLD *uk* and those that don't have the ccTLD *uk* should be the same as all the pages that include the term *pharmaceutical*, the numbers vary considerably for both Google and Live Search. In comparison to the figures of the other major search engines, Yahoo's difference of 30,000 out of over 56,000,000 hits seems very accurate.

Table 4-2 Discrepancies within search engine results

Query	Search engine results		
	Live Search	Google	Yahoo
Pharmaceutical site:uk	608,809	1,320,000	2,370,000
Pharmaceutical –site:uk	12,031,915	71,100,000	53,700,000
Total	12,640,724	72,420,000	56,070,000
Pharmaceutical	14,981,926	76,000,000	56,100,000

Despite the smaller index and the less accurate estimated results of the search engine, Live Search may still be considered the most appropriate for this investigation due to the additional operators available at the time of data collection. Although it would have been possible to utilise more than one search engine, e.g., Live Search for outlinks and Yahoo for inlinks, it was felt that it was more appropriate to use a single search engine so that comparisons could be made between inlinks and outlinks based on the same corpus of web pages, and so that reciprocal-links could be retrieved.

For each of the ABPI members' web sites in the investigation information was gathered from Live Search's API about the web site's inlinks, outlinks, reciprocal-links and the number of pages indexed by the search engine. Although Kretschmer et al. (2007, p.521) have said that the "weakest indicator of collaboration at the level of a hypertext would be a *reciprocal linkage* between two web sites", there is no guarantee as to whether a reciprocal-link will reflect collaboration, and conversely a uni-directional link has been shown to reflect collaboration on many occasions. It is therefore more appropriate to investigate both uni-directional links and reciprocal-links.

Web site inlinks were collected by combining the *linkdomain* and the *site* operator. The *linkdomain* operator finds pages that link to any page within a specified domain, whilst the *site* operator was used so that only external web links were retrieved.

e.g., linkdomain:wlv.ac.uk -site:wlv.ac.uk

Outlinks were collected using Live Search's unique *linkfromdomain* operator, again in conjunction with the *site* operator to prevent the inclusion of self-links. The *linkfromdomain* operator finds pages that are linked to from a specified domain.

e.g., linkfromdomain:wlv.ac.uk -site:wlv.ac.uk

As with the pilot study into university-industry-government relationships manifested through reciprocal-links, the MSN reciprocal-links were collected through combining the

linkfromdomain operator with the *linkdomain* operator. Combining the two operators finds those pages that are both linked to from a particular domain, and have a link to the particular domain (see section 3.6.2.1 for more details). As with the finding of inlinks and outlinks, the *site* operator was utilised to discount self-links.

e.g., linkfromdomain:wlv.ac.uk linkdomain:wlv.ac.uk -site:wlv.ac.uk

A program was written in visual basic to retrieve the inlinks, outlinks, and reciprocal-links for the identified web sites of each of the 145 organisations identified from the ABPI web site. Although the Live Search API only allows for the retrieval of 50 links at a time, it enables you to choose where those 50 links are taken from within the first thousand results. It is therefore possible to create a loop within the program and download up to the first thousand links.

The number of outlinks and reciprocal-links identified through the Live Search API not only depends on the links placed on the web site, but also on the extent that the web site has been indexed by the search engine. Therefore the *site* operator was used for each of the web sites to determine whether or not they had been extensively indexed by the search engine. Where there were not many pages indexed the URL was visited once again to determine the reason. An additional six URLs were identified that automatically forwarded browsers to a different URL, however, as Live Search had discontinued its *linkfromdomain* operator by this time it was no longer possible to collect the data, therefore these sites were excluded from the rest of the investigation.

4.3.3 Data cleaning

The link data collected from Live Search provides information about the links between web pages, and it is necessary for this data to be cleaned to provide information about the linking between organisations. This is necessary for determining the level of interlinking between the core web sites and for identifying the important partner organisations; a network of more than two nodes is unlikely to occur if the page is used as the node rather than an aggregation there of, as self-links are not included in the study. Whilst web pages may be aggregated in a number of ways, e.g., by domain, sub-domain, or directory, each is likely to combine some web sites which belong to different organisations, and divide parts of the same organisation. For example, aggregating according to sub-domain would treat links to jobs.bbc.co.uk and news.bbc.co.uk as links to different web sites, whilst links to the official MySpace page of Warwick University (www.myspace.com/warwickuniversity) and the official MySpace page of the English glamour model Jordan (www.myspace.com/Katie_2005) would be considered to be linking to the same web site.

Within this investigation the web pages are aggregated at the domain level. Many of the organisations listed on the ABPI web site are multinational corporations, and it seems likely that they and their collaborators will, for the most part, be using their own domain names rather than piggybacking on someone else's domain name.

4.3.4 Determining Live Search API coverage

Whilst it is recognised that search engines do not provide comprehensive coverage of the web (e.g., Snyder & Rosenbaum, 1999), there is no consensus about a suitable method for determining whether a search engine's coverage is sufficient for a particular webometric investigation. Although Bar-Ilan (2001) proposes that more than one search engine being used is liable to increase the coverage, it is still recognised that combining the results of all the main search engines will be far from exhaustive. Within this study it is proposed that the

suitability of a search engine's coverage is determined by investigating not only the number of links and web pages found for a particular set of queries, but that the distribution of these numbers is also used as a source of information about the sufficiency of a search engine's coverage. Previous investigations have found that many of the features of web sites follow Zipfian distributions when plotted in rank order, including the number of pages, the number of users, the number of inlinks, and the number of outlinks (Adamic & Huberman, 2001; 2002).

Whilst such findings were based on a large random sample of web sites, patterns would still be expected to emerge from a select group of web sites. Although adherence to a Zipfian distribution would not prove that the search engine has sufficient coverage of the web sites under investigation, it contributes to the reliability of any conclusions based on the data gathered, which can only ever be non-robust (Thelwall, 2004c). Equally wild divergence from a Zipfian distribution would not discount the validity of the search engine data, although it would indicate a need for closer inspection of the results, or even the need to use an alternative data collection method.

Using a distribution approach to investigating the suitability of the Live Search API is an attempt to establish a sustainable methodology rather than merely confirm the suitability of the Live Search API for one specific investigation. Although it would be possible to investigate the suitability of the data collection method through comparing the results found with the Live Search API with those obtainable with a web crawler, the findings would only show the suitability of the search engine to this particular study and would have risked disrupting people's web services. If future investigations could only determine the suitability of a search engine by recourse to crawling all the web sites in the study, then there would be no point in using the search engine data.

Within this investigation data is collected from the Live Search API about the number of inlinks, outlinks, MSN reciprocal-links, and indexed web pages for each of the ABPI members' web sites. This data is then plotted in rank order on log-log graphs with two trend lines: one based on a fit of least squares, and the other with ranked slope -1 based around the median ranked web site. The trend line of least squares enables the visualisation of how far the results differ from the straight line of a Zipfian distribution, whilst the trend line with a ranked slope of -1 provides a comparison based on the findings of previous investigations (Adamic & Huberman, 2001; 2002).

4.3.5 Hyperlink Network Analysis of the networks

Hyperlink network analysis (HNA), is the name given to the application of social network analysis (SNA) techniques to the web, taking hyperlinks to be representative of social and communicational ties (Park & Thelwall, 2003). Link analysis differs from HNA in that, rather than taking links to be representative of one thing, link analysis has tended to focus more on the validation process, with deeper investigation into possible interpretations of web links through classification of web links and correlations with external indicators. SNA analysis developed in parallel with scientometrics in the late 1970's (Leydesdorff, 2007), whilst more recently HNA has developed in parallel to webometrics.

The suitability of a social network analysis approach to informetric studies is well recognised and has formed the basis of numerous information science investigations (Otte & Rousseau, 2002), and as the web may be considered a social network of unprecedented scale (Otte & Rousseau, 2002), it is unsurprising to find that in recent years SNA has formed a part of some webometric investigations; looking at the network of relationships between actors

(e.g., Heimeriks et al., 2003; Kretschmer & Aguillo, 2004; Oretga & Aguillo, 2007), rather than the reasons for individual connections (e.g., Thelwall, Harries et al., 2003; Wilkinson, Harries et al., 2003; Bar-Ilan, 2004a).

In this study the SNA concept of centrality is applied to test the hypothesis that the central organisations in the web network are not necessarily those that are more collaborative, but rather those that are information providers. A central actor is one involved in many ties, connecting to other actors and being connected to from other actors, whilst a prestigious actor is one who is the object of many ties (Wasserman & Faust, 1994). Treating the network as non-directional, which would be the case if links represented collaboration, is necessary as otherwise the centrality of some of the links could not be calculated.

There are various different measures of centrality, each giving different emphasis to different aspects of the network. Within this investigation three of these centrality measures are used: degree centrality, closeness centrality, and betweenness centrality (defined below). These are the most established, and most important (Otte & Rousseau, 2002), of the measures of centrality, and may be simply calculated with Pajek network visualisation software. Whilst there are additional measures of centrality such as the eigenvector centrality and Kretschmer and Kretschmer's measure of weighted centrality (2006), these measures require non-binary networks (Park & Thelwall, 2003), whilst the search engine results are more appropriately treated as dichotomous networks. This is for two reasons. Firstly, due to the way the data is provided by Live Search, only providing one or two links per web site, also, if links do reflect collaboration, it does not necessarily mean that the number of links reflects the amount of collaboration.

Degree centrality is based on the number of direct links an actor has with other actors, and may be calculated for both directed and non-directed graphs (Wasserman & Faust, 1994). Degree centrality is the SNA equivalent of the crude counting of inlinks and outlinks within link analysis, and whilst it can provide useful information, it fails to take into consideration the position of the nodes within the network. Whilst an actor may have limited numbers of ties with other actors, those ties may be essential to the overall network. This issue addressed in the calculation of betweenness centrality, where an actor is considered central if it lies between other actors on their geodesics (Wasserman & Faust, 1994). Closeness centrality is based on the idea that an actor is central to a network if it can quickly interact with all other actors (Wasserman & Faust, 1994). Investigation of the closeness centrality means that the network is not overly influenced by the clusters of links in one area of the network.

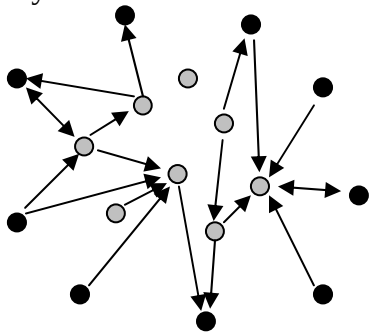

Figure 4-1 Illustrative network diagram of core and extended networks

Using the data collected through the Live Search API it is possible to investigate not only the network of web sites of the organisational members of the Association of the British

Pharmaceutical Industry, but also the extended network, including those web sites that link to, or are linked to from, the core ABPI members. However, there are differences between the relationships between the core web sites, and those web sites that are introduced in the extended network; this is illustrated in the network diagram in Figure 4-1. Within the network diagram the core web sites are shown as light grey dots, whereas those additional web sites identified by their linking with the core web sites are shown as black dots. Whilst the core web sites can be found to link with many other core web sites (or none) and many of the extended web sites (or none), the extended web sites must necessarily be linked with at least one core web site and the linking relationships between the extended web sites are unknown.

Within the study the degree centrality, the betweenness centrality, and the closeness centrality are calculated for both the core network and the extended network. Whilst it would seem likely that those organisations that have the highest centrality are all from the core set of organisations because the extended network must necessarily link to the core set and can not link to one another, the inclusion of the calculation of betweenness centrality and closeness centrality allows for extended network web sites to be included. Degree centrality is unlikely to include organisations from the extended network as whilst it is possible for a core web site to outlink to a thousand different web sites and have inlinks from a thousand different web sites, it is only possible for a non-core web site to be found to link to and from the 139 core web sites.

In addition to calculating the three types of centrality for both the core network and the extended network, the investigation considers the sites that are connected to the highest proportion of core network web sites, both through outlinks, inlinks, and reciprocal-links. This enables a comparison between the core sites with the extended web sites without either preferential treatment to the core sites, inherent in a network using the core sites as seeds, or allowing for the network to be heavily influenced by organisations whose contribution to the field is unknown.

To help determine the reason why some sites have either a higher degree of centrality, have more outlinks to the core network, or have more inlinks to the core network, a classification was carried out on a selection of the web links to and from these highly connected web sites. Utilising the same classification as was used to determine the reasons for link placement amongst the organisations as a whole (see Table 4-3), a sample of ten links were classified for each of the highly connected web sites in each of the categories.

4.3.6 Web link classification

As well as investigating those web sites that are most central to the networks as a way of investigating what web links reflect, the investigation also includes a classification of a sample of the web links from the total population. Previous investigations have shown that although there are a significant proportion of links that are placed that reflect a collaborative relationship, the majority reflect a non-collaborative relationship. These collaborations have focused primarily on academic web sites, whereas this investigation is primarily of organisations within the industry sector. Whilst it is expected that a higher proportion of MSN reciprocal-links will reflect a collaborative relationship than either inlinks or outlinks; it is necessary to carry out a classification of inlinks, outlinks and MSN reciprocal-links. The preliminary investigation into reasons for link placement on academic web sites (see section 3.4) has shown that there are difficulties in arriving at inter-classifier agreement for a fine grained classification of the reasons for link placement. This study therefore adopts the broad

classification utilised within the preliminary study of reciprocal-links between universities, industry and government organisations in the West Midlands (see section 3.6.3), classifying the relationships as either collaborative or non-collaborative.

The term *collaborative relationship* is applied to any relationship where the two organisations are thought to have a working relationship. This may include: research collaboration, one party carrying out work on behalf of another, one party providing sponsorship or funding, the two organisations working together towards mutual goals, as well as more informal relationships, such as speaking at a conference. The classification *non-collaborative relationship* is applied where the relationship is primarily for informational resources and references. The term 'collaborative relationship' is used in a broader sense than would be expected to be reflected within traditional co-authorship, allowing for the possibility of the web to provide information about the less formal relationships.

The classification category of *collaborative relationship* is further sub-divided according to whether or not the identified web site is an affiliated part of the original organisation, as many organisations use different web sites for different groups and services. A reciprocally-linked web site may be part of the same organisation (i.e., an affiliated web site), an external organisation (i.e., a non-affiliated web site), or a web site that is the result of a collaboration with another organisation. Thus an affiliated web site may actually belong to several collaborating organisations. Whilst the boundaries of an organisation are not always clear, in this study the boundary is determined by whether the web site contributes to the official work of the organisation. The term affiliated is used rather than subsidiary, which was utilised in the preliminary investigation (see section 3.6), as it seems more appropriate term for referring to separate parts of the same organisation where one part is not necessarily subordinate to the other.

The classification also takes into consideration that some links cannot be classified, either because of the web pages not being in English, or the page no longer being available, or the page having changed and the link no longer being identifiable. Although it is possible to investigate the reasons for the appearance of links that are no longer accessible on the current web page through analysis of a search engine's cached copy or through the internet archive, as was utilised within the preliminary investigation into the university to government web links (see section 3.3), with currency of the information being one of the attributes of the web that make it a potentially more suitable source of information than the traditional bibliometric surrogates, these were ignored.

Table 4-3 Classification of organisational relationship between the web sites

Type of relationship
1. Collaborative relationship (formal or informal)
a. Affiliated web site
b. Affiliated web site – with partners
c. Non-affiliated web site
2. Non-collaborative relationship
3. Unable to determine reason
a. Non-English web page
b. Link no longer available
c. Outlink source page could not be found

The classification of the relationship between web sites was carried out using content analysis techniques (Krippendorff, 2004), with both the source page and the target page analysed to see whether or not there was any indication of a collaborative relationship between the owners of the web pages. For the classification of the inlinks the source and target page were readily identifiable; Live Search provides the URLs of those pages which linked to a certain web site so it was possible to go to the source of the link, and then on that page identify the target page as the domain was already known. However, for the classification of the outlinks it was necessary to search for the specific source page as when using the *linkfromdomain* operator Live Search provides a list of those sites that the web site links to, not the source page the links were on. During the short period of time between the data collection and the link classification Live Search withdrew the *linkdomain* command and it was therefore no longer possible to search the Live Search database to find the source page; instead Yahoo's database was used, with an additional sub-section added to the 'Unable to determine reason' category, where the source page could not be found (see Table 4-3). When the same page was found to be linked to by more than one page in the source web site, one page was selected at random to be the source page investigated. When searching Live Search for MSN reciprocal-links it returns a list of those web pages that have a link back to the original web site, the source page that was analysed was the one found by the search engine, whilst the target page was the one identified by the link in the source page. Only two web pages were analysed to determine whether each link represented a collaborative relationship, as otherwise the increased number of web pages may account for increased classification of collaboration.

Whilst it is possible that there is a relationship between two linked organisations that is not immediately apparent from the web site, no attempt was made to determine whether such a relationship can be identified in other primary sources in the manner of Vasileiadou and van den Besselaar (2006). It is highly likely that global organisations involved within the same industry sector collaborate with each other at some level and that this information will be obtainable from other primary sources. However, the purpose of this study is to determine why links are being placed and what can be inferred from them, therefore a content analysis approach is far more appropriate.

As with previous extensive classifications, for each of the three link sets (inlinks, outlinks, and reciprocal-links), ten links were taken from the search engine results for each of the web sites in the study, so that the sample was not biased towards the organisations with a greater number of links. A random selection of 200 web links was then selected from this sample for classification, of which a second classifier classified 20%. Classifier reliability was determined by calculating Krippendorff's alpha coefficient (Krippendorff, 2004); this takes into account the chance agreement, the magnitude of the misses, and adjusts for whether the variable is nominal, ordinal, interval, or ratio (Neuendorf, 2002).

4.3.7 Traditional bibliographic data collection

For each of the 139 ABPI organisational members with identifiable web sites data was also collected about their scientific publications and patent applications so that comparisons could be made between the information available on the web and traditional bibliometric sources. The information about scientific publications was collected from the Thomson ISI Web of Science and the patent information from the US Patent and Trademark Office. The majority of previous bibliometric investigations of scientific articles have been operationalized through the Thomson ISI Web of Science (WoS) or its predecessors, and although a number of

competitors have emerged in recent years (Roth, 2005), which have been the focus of many recent comparative studies (e.g., Jacso, 2005; Norris & Oppenheim, 2007), the WoS is still considered one of the most comprehensive databases. In addition, as the traditional database of choice, further tools have been developed by bibliometric researchers to manipulate the WoS bibliographic files once they have been downloaded from the web site. The US Patent and Trademark Office (USPTO) database was chosen rather than the databases of the UK Intellectual Property Office (UKIPO), or the European Patent Office (EPO), due to the searching facilities provided by the database. Patent databases contain hundreds of years of records and, as many of the organisations under investigation have long histories, it is necessary to be able to limit the search according to certain years. Whilst the EPO and the UKIPO allow for searching by a specific date, they do not at the time of investigation allow for the searching by a range of dates. This problem is further complicated by the results returned in the order that they were uploaded to the database rather than the order the patents were issued, and with only the first 500 patents viewable there are likely to be many pertinent patents that are not viewable. As the pharmaceutical industry is very much a global industry, and the US market is a major player, it seems reasonable to suppose that the patents applied for from the UKIPO will also be applied for in the USPTO.

Within both databases an organisation's documents were identified through searching for a truncated version of the organisational name: within the assignee name field in the USPTO; and the organisational name field within the WoS. The truncated version of the organisational name allows for different divisions of the organisation to be retrieved. If, however, the search retrieved a disproportionately large number of patents from a similarly named organisation, which was not thought to be part of the organisation under investigation, a truncated version of the organisational qualifier, e.g., uni, was incorporated along with a wildcard to allow for variations. The assignee name contains the name of "the individual or entity to whom ownership of the patent was assigned at the time of patent issue" (USPTO, 2007), and may contain one or more organisations. Equally the organisational name field in the WoS may contain more than one organisation. Within this study organisational collaboration was inferred when two different organisations were found within either the assignee name field of a patent or the organisational name field of a scientific publication.

Both the USPTO and the WoS have bibliographic records going back over many decades, whereas the web has been available for a far shorter length of time and is more ephemeral in nature. It is therefore necessary to restrict those traditional bibliographic surrogates that are investigated within this study to the more recently published patents and scientific publications. The USPTO has both a database of issued patents and a database of patent applications. Analysis of issued patents incorporates a substantial delay due to the time it takes for patents to be issued, and it is therefore recommended that patent applications are analysed, which are published by USPTO 18 months after their priority date (Hinze & Smooch, 2004). Whilst analyses of patent applications necessarily includes analysis of applications for patents that will not be granted, this is immaterial to an investigation of collaboration between organisations; whether or not a patent is issued, the application may still be seen as evidence of collaboration. Those patent applications published in the previous year were investigated, as a year is thought to be a reasonable period of time within which to expect a collaboration to apply for a patent and the date recent enough to be comparable with the link data. The WoS does not allow the retrieval of the records between specific dates, but rather allows the retrieval of records for particular years; therefore the records for those articles published in both 2006 and 2007 were investigated.

Each of the patent applications was manually examined to determine whether the record related to the organisation under investigation, and if so, whether there were any other organisations included within the assignee name field. 'Other' organisations were defined as those that had sufficiently different organisational names that a connection was not immediately apparent between the two. Although many patents have more than one assignee name it is not unusual for them to be separate divisions of the same organisation. The WoS records were downloaded (500 at a time), and BibExcel was used to extract the organisational field names. Due to the far larger number of WoS records found than USPTO records found, the data was not browsed at this stage, but rather checked on an individual basis as necessary.

4.3.7.1 Collaborative relationships not visible through traditional bibliometric sources

One of the original premises of this investigation into the web manifestations of knowledge-based innovation systems, is the idea that the web can provide insights into relationships between organisations not visible within traditional bibliometrics. This is tested in this investigation through exploring whether those collaborative relationships that are reflected through web links are also visible through the traditional bibliometric sources of patents and science articles. Each of the web sites that were classified as having a non-affiliated collaborative relationship was investigated to determine whether a relationship between the represented organisations was visible in either the WoS or the USPTO. The comparison also enables the testing of one of the conclusions of the extensive classification of university outlinks to the different domains: that there may be a large number of web links that reflect a collaboration, but that it only becomes clear that web links reflect collaborations when the collaborations are already known as there is insufficient information on a web page (see section 3.4).

Each of the pairs of web sites that were classified as having a non-affiliated collaborative relationship comprises of one web site belonging to an organisation that is part of the original set of ABPI organisations, i.e., the core web site, and one organisation that may or may not be one of the ABPI set of organisations, i.e., the linked web site. Although the organisational name is known for the core web site, for most of the linked web sites it was necessary to identify the organisation they represented. This was achieved through analysis of the linked page as well as the homepage of the web site that contained the linked page. The organisational name was not necessarily the dominant name on the web site, but rather was the copyright owner identified at the bottom of the page.

Both the downloaded WoS records and USPTO records were then investigated to see if the seed organisation had a collaborative relationship with the linked organisation.

4.3.7.2 Collaborative relationships not visible through web links

As well as investigating whether the relationships visible through web links are visible through traditional bibliometric data sources, it is equally necessary to investigate whether there are collaborative relationships visible within traditional bibliometric sources that are visible through web links; whether collaboration visible through web links can be used to supplement traditional bibliometrics, or to supplant it.

A random selection of 50 relationships visible within the organisational name field of the identified scientific articles, and 50 relationships visible within the assignee name field of the identified patents were investigated to see whether they were visible within the web links that were downloaded through the Live Search API. As with the classification of the web

links, to prevent bias towards the larger organisations with many organisational relationships, a sample of ten organisational relationships were selected from each of the core organisations to produce two lists, one of technological collaboration and one of scientific collaboration. The 50 random relationships for investigation were then taken from each of these sample lists.

As many of the organisations have multiple web sites it is impossible to determine with 100% accuracy whether a web site belongs to an organisation without visiting the web site, and even then there is no guarantee that there will be the details necessary to make the connection with the original organisation. With many of the seed organisations identifying hundreds of linked web sites it is not a practical solution to view each of the web sites. Firstly, the name of the organisation was searched for within a number of search engines to find an organisation's official web site. The domain name of the web site was then extracted from the URL and searched for within the lists of inlinked, outlinked and reciprocally-linked web sites to see if it occurred in connection with the core web site. It was searched for without the second/TLD domain to allow for the identification of certain affiliated web sites that may have a different ccTLD. The identified URLs were then visited to confirm whether or not they were related.

5 Results

5.1 Live Search coverage

As would be expected there was a wide variety in the number of pages indexed by Live Search, as well as between the numbers of inlinks, outlinks and MSN reciprocal-links for each of the web sites (see Table 5-1).

Table 5-1 Estimated number of web pages and links for core web sites

	Pages Indexed in Live Search	Inlinks	Outlinks	MSN Reciprocal-links
Lowest	1	2	0	0
Highest	17,033,697	32,040,380	881,247	105,588
Mean	127,268	237,756	6,391	766
Median	1037	202	13	1

The estimated number of pages indexed, number of outlinks, number of inlinks, and number of reciprocal-links for each of the web sites are shown in rank order in log-log graphs in Figure 5-1, 5-2, 5-4, 5-5. The trend-line for the results (shown with a solid black line) is based on a fit of least squares, a dashed black line shows how the results would differ if Microsoft was not included in the results, whilst a ranked slope of -1 based around the median ranked web site is also shown on the graphs with a dotted black line so that a comparison can be made with the expected results for a random sample of pages from the web (Adamin & Huberman, 2001; 2002). With the exception of the Microsoft web site, which is ranked first for each of the four criteria, the other highly ranked web sites do not have quite as many pages indexed, outlinks, inlinks, or MSN reciprocal-links as may be expected by the trend-line, whilst the lower ranked web sites fail to have as many inlinks, outlinks, and pages indexed as expected.

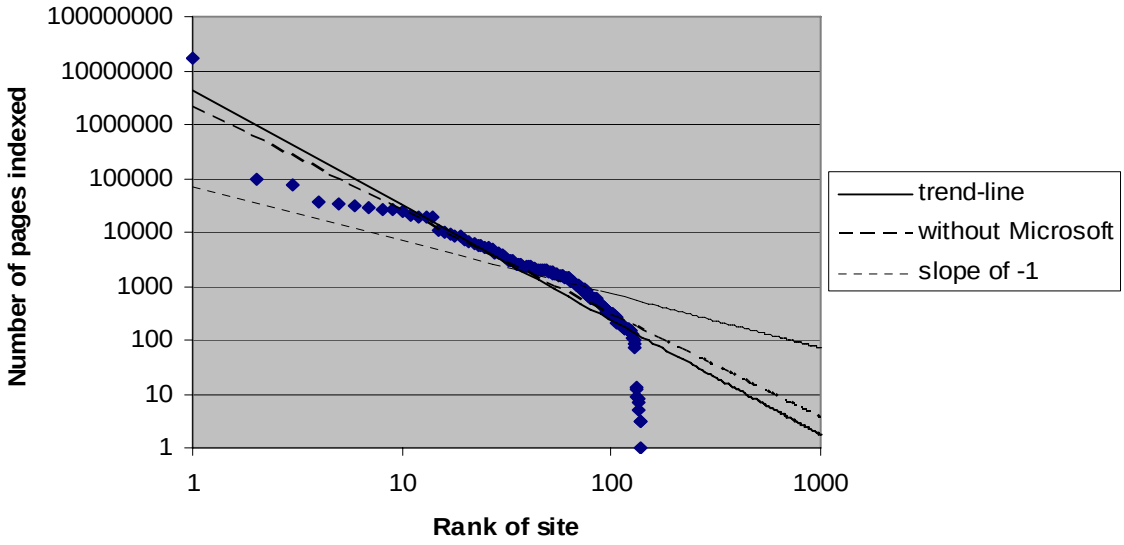

Figure 5-1 Web sites rank ordered by the estimate number of pages indexed

Figure 5-1 shows the estimated number of pages indexed for each of the web sites in the investigation in rank order on a log-log axis. The graph is not linear, with more Microsoft pages being reported indexed than would be expected, whilst the lower ranked web sites have far fewer pages reported indexed than would be expected from a perfect Zipf law fit. Closer investigation of those sites with less than ten pages reported indexed finds that four are old web sites that no longer seem to be updated by the organisations, two of the web sites only contain a few pages of links pointing to additional affiliated organisational web sites, whilst one of the web sites is quite extensive, but is in Flash rather than the more usual search engine accessible HTML.

Figure 5-1 also contains a noticeable gap in-between the 130th and the 131st ranked web sites, where the number of pages in the index drops from 75 to 13. A similar sudden drop is also visible in Figure 5-2.

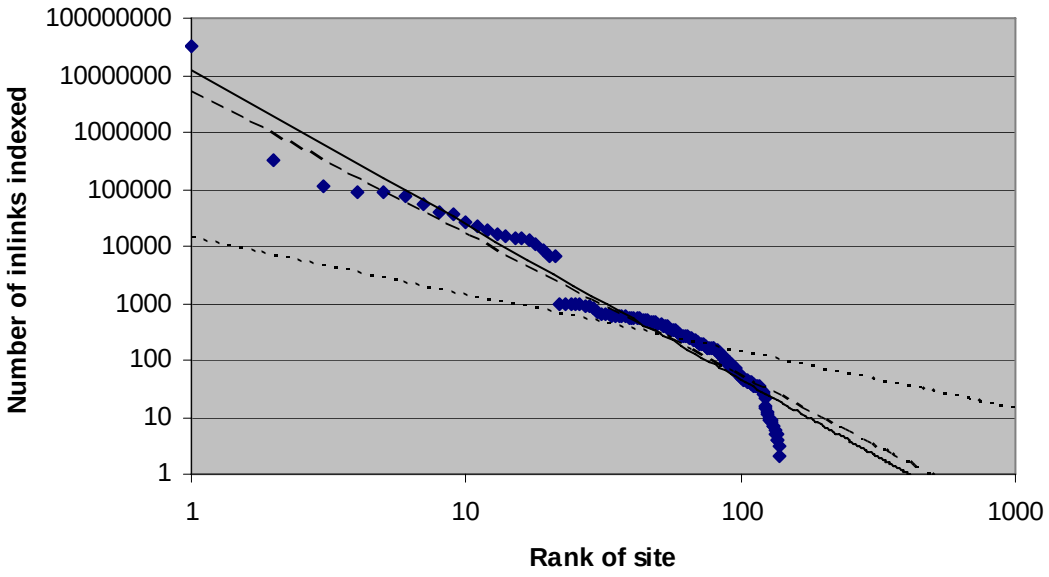

Figure 5-2 Web sites rank ordered by their number of inlinks

An explanation for the gap in the results may be found in the final pilot study (see section 3.6) where it was established that the estimated number of MSN reciprocal-links is not a static number, but rather varies according to how deep the results are retrieved from, apparently becoming more accurate the deeper the results are retrieved from. This discrepancy between the estimated number of links and actual number of links is clearly visible in Figure 5-2, where there is a discernable gap between the 21st ranked web site which has an estimated 6, 420 inlinks and the 22nd ranked web site which has an estimated 975 indexed inlinks.

Even though the web sites that have less web pages indexed do not generally have as many inlinks, they do nonetheless have some, and on many occasions may be thought to punch above their weight (see Figure 5-3).

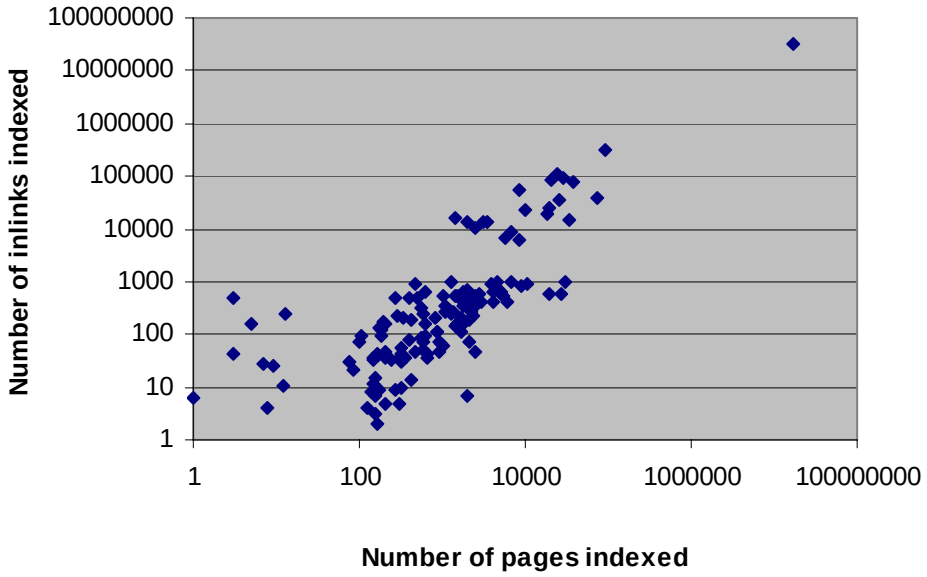

Figure 5-3 Scatter graph showing the number of inlinks compared with the number of pages indexed

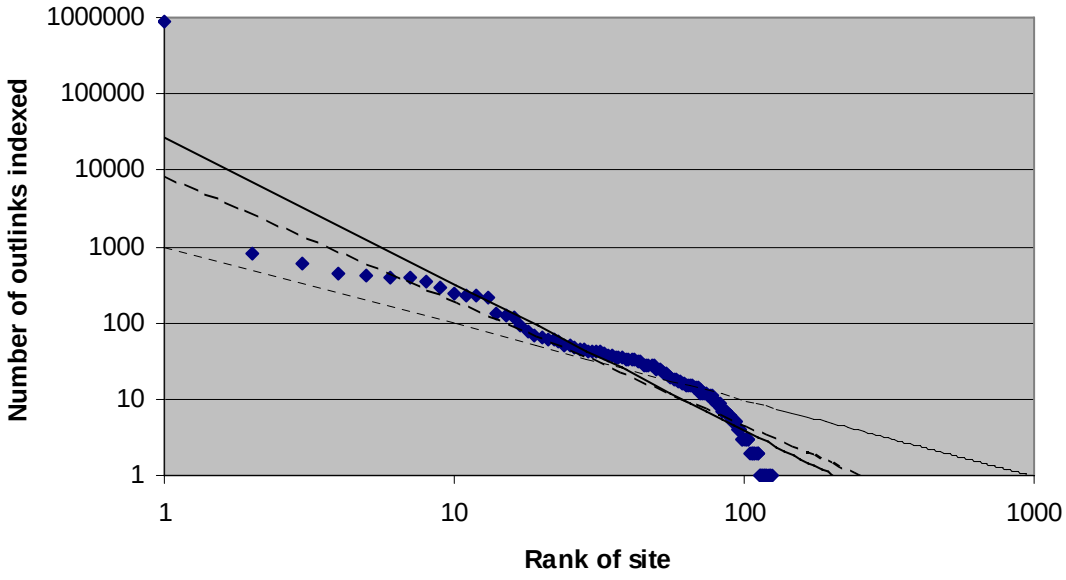

Figure 5-4 Web sites rank ordered by their number of outlinks indexed

Unsurprisingly for a set of organisations that were mostly from the industry sector, there were far more inlinks than outlinks, with only three organisations having more outlinks than inlinks, and only one organisation had more than 1,000 outlinks indexed by Live Search: Microsoft. Microsoft was also the only organisation to have more than 1,000 reciprocal-links indexed. 21 of the organisations had no outlinks, and these are necessarily missing from the graph.

The over-estimation of the number of results can be seen as a contributory factor to the differences between the trend lines; those graphs where many of the sites have more than 1,000 pages indexed have a higher gradient. Figure 5-2, web sites rank ordered by their number of inlinks, has the highest gradient at -2.72 and 21 web sites that are found to have over 1,000 inlinks, whereas the gradients for the web sites rank ordered by their number of outlinks and reciprocal-links are only -1.91 and -1.78 respectively. It is worth noting that there is not the same gap in Figure 5-1 which shows the estimated number of pages indexed, or as steep a gradient as for the number of inlinks, despite over 71 of the sites having more than a thousand pages indexed. This is due to the fact that the estimated number of pages indexed was based on running a single query, rather than digging down to retrieve the results, and as such it is only the 15 web sites that have around ten or less pages indexed which are out of sync with the rest.

Figure 5-5 shows the web sites rank ordered by their number of reciprocal-links, this is found to have the least difference between the two trend lines; although the graph also excludes the most web sites, as only 75 of the 139 web sites were found to have any MSN reciprocal-links indexed.

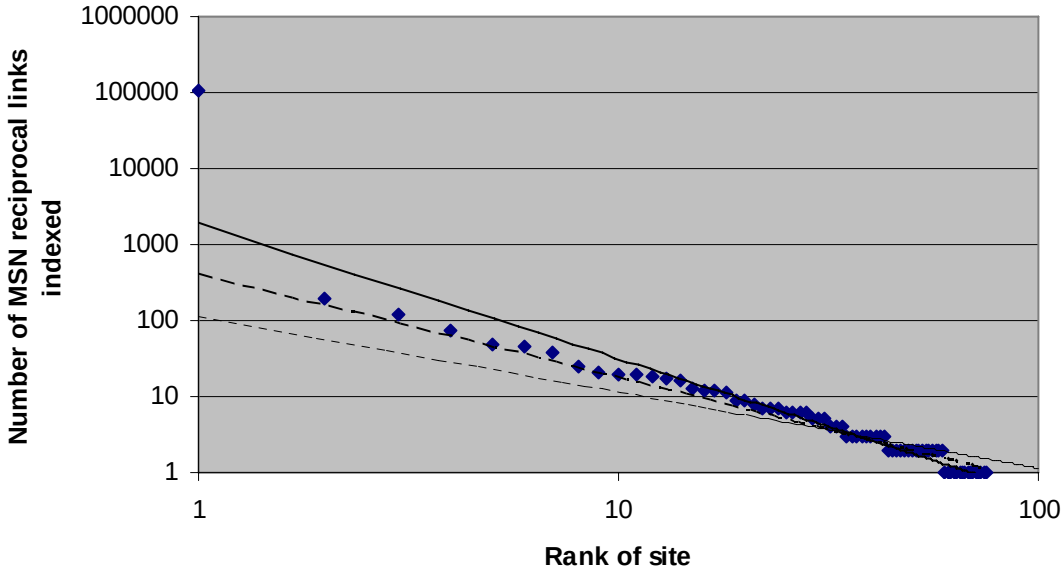

Figure 5-5 Web sites rank ordered by their number of MSN reciprocal-links

5.2 Linking between the core organisational web sites

Despite the study including 139 organisations that were all members of the same organisational body, there was relatively little interlinking between the core organisations. Whilst it was possible for each of the 139 organisations to link to each of the other 138 organisations in the study, totalling 19,182 possible directional links, very few of these were found. Only 38 outlinks were found from any one of the core web sites to another of the core web sites, 57 inlinks were found to any one core web site from another core web site, and two MSN reciprocal-links were found to exist between the core web sites. Although theoretically the number of inlinks should equal the number of outlinks, the discrepancy between the numbers, as well as the actual links the numbers represent, can be attributed in part to the

retrieval of only the first thousand inlinks, outlinks and MSN reciprocal-links from each of the organisations. 21 of the core web sites had more than a 1,000 inlinks, including Microsoft which also had thousands of outlinks and thousands of MSN reciprocal-links. It is possible that there were inlinks for each of these 21 web sites from other core web sites, which were not recognised because they were not ranked in the top 1,000. This is most apparent when looking at the interlinking of Microsoft: only one of the core organisations was found to be linking to Microsoft when examining Microsoft's inlinks, whereas 20 web sites were found to be linking to Microsoft when examining the other core web sites' outlinks. With over 32 million pages linking to Microsoft, it is not surprising to find that over a thousand are ranked higher than some of the core organisations in this study.

Combining the data gathered about the core web site's inlinks, outlinks, and MSN reciprocal-links finds just 81 relationships out of a possible 19,182. Altogether 123 of the 139 organisations don't link to any of the other core web sites, whilst 113 of the organisations aren't linked to by any of the other core web sites. 64 of the core web sites have a link of some sort, i.e., at least one inlink or one outlink. These links produce one large 53-node network (see Figure 5-6), as well as four two-node clusters and a three-node cluster. Figure 5-6 shows a directed graph of the interlinked web sites, the positions of the nodes are based on the Kamada and Kawai's (1989) algorithm, which tries to position connected sites close together whilst ignoring the strength of the connection (Leydesdorff & Vaughan, 2006).

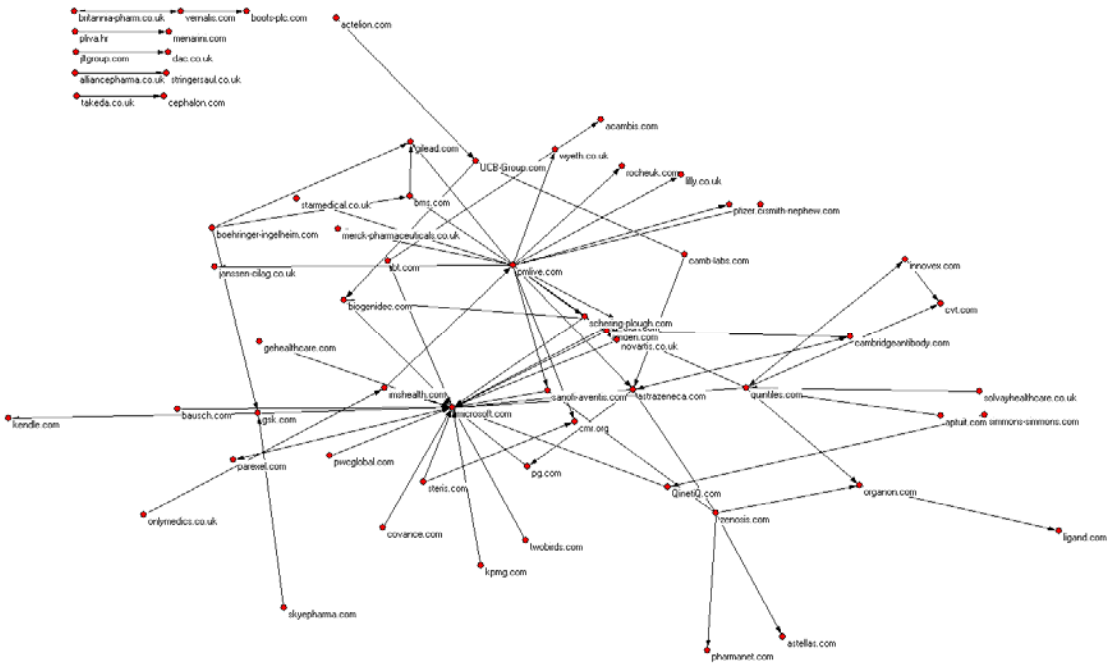

Figure 5-6 Fifty-three-node network of interlinked core web sites

Almost half of the relationships involve either one of two organisations, either Pharmaceutical Marketing Ltd, for which 17 outlinks and 1 inlink were found, or Microsoft, for which 20 inlinks were found and two outlinks. Pharmaceutical Marketing Ltd is a publishing company and pmlive.com, its web site, provides a range of brief articles about the industry, a directory of organisations in the industry, as well as a jobs section for organisations in the industry. It is surprising that it only links to 17 of the organisations within the core group. Equally, it may be thought surprising that only 20 organisations linked to the Microsoft web site. These links

were placed for highlighting a piece of software that was required for viewing rich content, e.g., Word Viewer and Windows Media Player, or for highlighting other information on the vast Microsoft web site. Without the inclusion of these two sites, which are not found to be reflecting collaborative relationships, the largest connected network would only consist of only twenty web sites.

The degree centrality, betweenness centrality, and closeness centrality were calculated for each of the nodes using the Pajek network visualisation software. The nodes with the highest degree of centrality for each of the three sorts are shown in Table 5-2. Eleven of the web sites are ranked joint seventh with the same degree centrality; each has three ties with other organisations.

Table 5-2 The top ten nodes with the highest level of centrality amongst the core network

Rank	Degree Centrality		Betweenness Centrality		Closeness Centrality	
1	microsoft.com	0.403846	microsoft.com	0.593866	microsoft.com	0.536083
2	pmlive.com	0.346154	pmlive.com	0.384213	astrazeneca.com	0.472727
3	quintiles.com	0.134615	quintiles.com	0.193766	pmlive.com	0.448276
4	astrazeneca.com	0.115385	astrazeneca.com	0.147583	sanofi-aventis.com	0.444444
5	zenosis.com	0.096154	gsk.com	0.099799	schering-plough.com	0.440678
6	gsk.com	0.076923	zenosis.com	0.096933	quintiles.com	0.429752
7	UCB-Group.com, amgen.com, biogenidec.com, elan.com, bms.com, gilead.com, boehringer-ingelheim.com, imshealth.com, organon.com, sanofi-aventis.com, schering-plough.com	0.057692	schering-plough.com	0.065431	elan.com	0.42623
8			sanofi-aventis.com	0.062339	amgen.com	0.422764
9			biogenidec.com	0.055556	imshealth.com	0.422764
10			imshealth.com	0.054622	novartis.co.uk	0.416

5.3 Linking amongst the extended network: Additional important partner organisations identified from different sectors

For the full dataset operationalizing the web site definitions on a domain basis and combining the inlinks, outlinks and MSN reciprocal-links produces a graph of 29,293 nodes, of which there is a cluster of 29,289 web sites, and one of four web sites.

5.3.1 Web sites with high centrality

Within the extended network calculating the centrality of the nodes finds a change in those web sites with the highest centrality (see Table 5-3). Of the ten web sites with the highest degree centrality in the extended network, seven were not amongst those with the highest degree centrality in the core network, only Microsoft, GlaxoSmithKline and Bristol-Myers Squibb remain. Out of the ten web sites with the highest betweenness centrality in the extended network only two are amongst the top ten betweenness centrality for the core

network, i.e., Microsoft and GlaxoSmithKline, whilst the Association of the British Pharmaceutical Industry's web site shows the central role of an organisation outside of the core set to the network. Amongst the ten web sites with the highest closeness centrality, only one web site remains from the core network, Microsoft. All the other identified web sites are external to the core network.

Table 5-3 The top ten nodes with the highest level of centrality amongst the extended network

Rank	Degree Centrality		Betweenness Centrality		Closeness Centrality	
1	.microsoft.com	0.083003	.microsoft.com	0.154368	.abpi.org.uk	0.479189
2	.bt.com	0.044558	.bt.com	0.073068	.blogspot.com	0.451857
3	.gehealthcare.com	0.036875	.abpi.org.uk	0.065709	.yahoo.com	0.428663
4	.pg.com	0.036261	.gehealthcare.com	0.059303	.microsoft.com	0.417113
5	.pwcglobal.com	0.036192	.pg.com	0.058201	.google.com	0.409559
6	.gsk.com	0.034724	.pwcglobal.com	0.055548	.acs.org	0.385247
7	.bms.com	0.034349	.smith-nephew.com	0.053205	.stanford.edu	0.385034
8	.smith-nephew.com	0.034144	.kpmg.com	0.052249	.live.com	0.383717
9	.kpmg.com	0.033632	.bausch.com	0.051602	.bbc.co.uk	0.382995
10	.bausch.com	0.031344	.gsk.com	0.047251	.google.co.uk	0.377374

A classification of the reasons for link placement was carried out for each of the web sites that were included amongst the top ten highest levels of centrality; resulting in the classification of links to nineteen different web sites. Ten were from the core network (see Table 5-4), and nine from the external network (see Table 5-5).

Table 5-4 Classification of reasons for link placement with the most central core web sites

	.microsoft.com	.bt.com	.gehealthcare.com	.pg.com	.pwcglobal.com	.gsk.com	.bms.com	.smith-nephew.com	.kpmg.com	.bausch.com
Collaborative relationship										
-Affiliated web site	1	1	4	4		3	2		2	
-Affiliated web site with partners										
-Non-affiliated web site	3	2	4	2	3	5	4	5	6	2
Non-collaborative relationship	**3**	**6**	**0**	**1**	**5**	**1**	**4**	**2**	**2**	**5**
Page not available										
-Non-English web page				2				3		3
-Link/page no-longer available			1			1				
-Outlink source page not identified	3	1	1	1	2					

The most noticeable thing about the classification of reasons for link placement for the web sites with the highest centrality is that whilst those from the core network have links placed for both collaborative and non-collaborative reasons, the links with the web sites with high centrality in the extended network are overwhelmingly for non-collaborative reasons. The exceptions are the .abpi.org.uk web site, which may be considered an anomaly as membership

of the ABPI was a prerequisite for inclusion in the study, and standford.edu. No conclusion may be drawn about live.com because few of its pages are in English.

Table 5-5 Classification of reasons for link placement with the most central web sites not in the core network

	.abpi.org.uk	.blogspot.com	.yahoo.com	.google.com	.acs.org	.stanford.edu	.live.com	.bbc.co.uk	.google.co.uk
Collaborative relationship									
-Affiliated web site							1		
-Affiliated web site with partners			1						
-Non-affiliated web site	10					5	1		
Non-collaborative relationship		**10**	**7**	**10**	**10**	**5**		**10**	**10**
Page not available									
-Non-English web page							8		
-Link/page no-longer available			2						
-Outlink source page not identified									

5.3.2 Web sites linking to the core network

Table 5-6 shows the domains that link to the highest proportion of the core web sites. As selection of the core web sites was based on the listed members of the ABPI, and most of the organisational details included a link to the organisational web site, it is not surprising to find the ABPI's web site to link to the most core web sites, 126 out of 139. In addition to including further significant web sites within the pharmaceutical industry, the other domains that have links to many of the core web sites include search engines and directories (e.g., google.com and dmoz.org), blog and web site hosting domains (e.g., blogspot.com and geocities.com). Of the top eleven web sites only three may be considered subject specific at the domain level.

Table 5-6 Top 11 inlinking domains

	URL	Proportion of 145 Core Sites Linked to
1	.abpi.org.uk	126
2	.blogspot.com	72
3	.google.com	71
4	.answers.com	57
5	.yahoo.com	57
6	.unimotion.co.uk	52
7	.drugsontrial.com	50
8	.acs.org	49
9	.geocities.com	47
10	.dmoz.org	43
	.google.co.uk	43

All of the top eleven inlinking domains are part of the extended network of web sites, rather than the original core set, and whilst six of the web sites appeared within the list of web sites with the highest closeness centrality, there are an additional five web sites that didn't: answers.com, unimotion.co.uk, drugsontrial.com, geocities.com, and dmoz.org. The classification of the reasons for link placement can be seen in Table 5-7.

Table 5-7 Classification of the reasons for link placement on the top 11 inlinking domains

	.abpi.org.uk	.blogspot.com	.google.com	.answers.com	.yahoo.com	.unimotion.co.uk	.drugsontrial.com	.acs.org	.geocities.com	.dmoz.org	.google.co.uk
Collaborative relationship											
-Affiliate web site											
-Affiliate web site with partners											
-Non-affiliate web site	10							1	1		
Non-collaborative relationship		**9**	**10**	**10**	**7**	**10**	**10**	**9**	**8**	**7**	**10**
Page not available											
-Non-English web page										3	
-Link/page no-longer available		1			3				1		
-Outlink source page not identified											

As with the external web sites with a high centrality (see Table 5-5), those web sites that are connected to the most core web sites have links reflecting non-collaborative relationships. The exception, once again, is abpi.org.uk.

5.3.3 Web sites highly linked to from the core web sites

Table 5-8 shows the 12 web sites that are linked to by the most core web sites, with three web sites holding joint tenth position. The collection of web sites is not limited to the pharmaceutical sector, but rather includes a cross-section of search engines, web site enhancing tools, and a news site, as did the inlinking web sites. Unlike the list of inlinked web sites, the top 12 most linked-to web sites include five government web sites.

With the exception of Microsoft.com, all the top linked domains are outside the core network, and of these only Yahoo.com was identified as important through calculating the centrality of the domains for the extended network. Yahoo.com was also the only web site that was both amongst the top inlinking domains and outlinked domains. The classification of the reasons for link placement can be seen in Table 5-9.

Table 5-8 Top 12 Outlinked domains

	URL	Number of core sites that link to the domain
1	.adobe.com	39
2	.corporate-ir.net	23
3	.microsoft.com	20
4	.europa.eu	15
5	.sec.gov	14
6	.fda.gov	13
7	.nih.org	13
8	.medicines.org.uk	11
9	.macromedia.com	11
10	.europa.eu.int	10
	.real.com	10
	.yahoo.com	10

Once again there is little ambiguity amongst the reasons for link placement; with the exception of .corporate-ir.net, the web links reflect non-collaborative relationships.

Table 5-9 Classification of the reasons for link placement to the top 12 outlinked domains

	.adobe.com	.corporate-ir.net	.microsoft.com	.europa.eu	.sec.gov	.fda.gov	.nih.gov	.medicines.org.uk	,macromedia.com	.europa.eu.int	.real.com	.yahoo.com
Collaborative relationship												
-Affiliate web site		9										
-Affiliate web site with partners												
-Non-affiliate web site												1
Non-collaborative relationship	**7**		**6**	**3**	**5**	**10**	**3**	**5**	**5**	**4**	**8**	**4**
Page not available												
-Non-English web page							2		2	2		
-Link/page no-longer available			1						1			
-Outlink source page not identified	3	1	3	7	5		5	5	2	4	2	5

5.3.4 Web sites with more than one reciprocal-link

Table 5-10 shows those web sites that had more than one reciprocal-link with one of the core web sites. There were 16 web sites that were found to be reciprocally linked to more than one of the core set of web sites, but no site was found to we reciprocally linked to by more than two.

Table 5-10 Top 16 reciprocally linked domains

	URL	Number of 145 Core Sites Linked-to
1	.uk.com	2
2	.acs.com	2
3	.cams.ac.uk	2
4	.msn.com	2
5	.tysabri.com	2
6	.webhire.com	2
7	.asp.net	2
8	.cantos.com	2
9	.shareholder.com	2
10	.medicines.org.uk	2
11	.vcall.com	2
12	.pharmiweb.com	2
13	.live.com	2
14	.msdn.com	2
15	.hemscott.com	2
16	.sourceforge.net.com	2

With no web sites found to be reciprocally linked with more than two of the core web sites it was not useful to classify the links of the most highly reciprocally linked web sites.

5.4 A higher proportion of reciprocal-links reflect collaborative relationships than inlinks or outlinks

The primary classifier classified 200 inlinks, outlinks, and MSN reciprocal-links, and Table 5-11 shows the classification results. A second classifier classified 40 inlinks, outlinks, and reciprocal-links. Krippendoff's alpha coefficient of inter-classifier agreement was calculated as 0.88 for those relationships that could be classified. Although there is no single acceptable level of reliability, a Krippendorff's alpha of over 0.8 is generally considered acceptable for most investigations (Krippendorff, 2004). Due to various proportions of the inlinks, outlinks, and reciprocal-links being classifiable, the percentage of classifiable links represented by the collaborative and non-collaborative relationship categories are also provided.

Table 5-11 Classification of web links

	Inlinks	Outlinks	MSN Reciprocal-Links
Collaborative relationship	**62 (40.8%)**	**82 (66.1%)**	**133 (92.4%)**
-Affiliate web site	17 (11.2%)	52 (41.9%)	115 (79.9%)
-Affiliate web site with partners	0	2 (1.6%)	2 (1.4%)
-Non-affiliate web site	45 (29.6%)	28 (22.6%)	16 (11.1%)
Non-collaborative relationship	**90 (59.2%)**	**42 (33.9%)**	**11 (7.6%)**
Page not available	**48**	**19**	**56**
-Non-English web page	27	8	42
-Link/page no-longer available	21	11	14
-Outlink source page not identified	0	57	0

The classification finds that a relatively high proportion of inlinks, outlinks, and MSN reciprocal-links are found to reflect a collaborative relationship, and the 66.1% of outlinks reflecting a collaborative relationship is substantially higher than the classification of academic web links in earlier investigations (see sections 3.3 and 3.4).

5.5 Visibility of web identified collaboration in patents and science articles

Those organisations that were identified as having a collaborative relationship with a non-affiliated web site, either through an inlink, an outlink, or a reciprocal-link, were searched for in the assignee name field of the records from the USPTO patent applications database and the organisational address of the records from the WoS database. As can be seen from Table 5-12 very few of the collaborative relationships identified within the web were visible within the traditional bibliometric surrogates.

Table 5-12 Collaborative web links represented by patents and science articles

	Inlinks	Outlinks	MSN Reciprocal-Links
Number of collaborative relationships with non-affiliate web sites	45	28	16
Reflected in patents	0	0	0
Reflected in science articles	1	2	0

5.6 Visibility of patent and science paper identified articles in web links

A total of 50 organisational collaborations that were identified through the USPTO and 50 organisational collaborations that were identified through the WoS database were investigated to see whether the relationships were reflected in the web links. The results can be seen in Table 5-13. Few of these relationships were visible in the web links retrieved with Live Search, and these were only reflected in site inlinks.

Table 5-13 Collaborative patents and science articles represented by web links

	Additional web site not identified	Visible in Inlinks	Visible in outlinks	Visible in MSN reciprocal-links
USPTO identified collaboration	4	2	0	0
WoS identified collaboration	4	4	0	0

The six organisational collaborations that were also visible in either a patent or a science article were investigated to see whether the collaboration would have been identified from the web links alone. Of the two patent identified relationships, one would have been identified from the web link, whilst one was in Japanese. Of the four science article identified relationships, a collaborative relationship would have been identified for two of them, one was in Croatian, and one was a news story that would not have been identified as expressing a collaborative relationship between the two organisations.

6 Discussion

6.1 Introduction

This chapter discusses the results of the link analysis of the members of the Association of the British Pharmaceutical Industry, both specifically and within the broader context of the thesis more generally. The bulk of the chapter is split into five sections:

- Investigating collaborative relationships with a search engine
- Investigating the reasons for link placement
- The differences between webometric and traditional sources of information
- The potential of the web to provide collaborative information beyond the UK's pharmaceutical organisations
- A proposed methodology for using web links as a source of information about organisational collaboration in the future.

6.2 Investigating collaborative relationships with a search engine

Despite the recognised limitations of using search engines for data collection in webometric studies (e.g., Snyder & Rosenbaum, 1999), a search engine was successfully used within the link analysis of the members of the ABPI, as well as within the two pilot investigations (see sections 3.5 and 3.6), to find information about inter-organisational collaboration. This section discusses the limitations of Live Search, the advantages of utilising the distribution of a search engine's results in addition to the raw numbers, and why, despite the limitations, search engines continue to offer the best solution to the problem of data collection when investigating collaborative relationships between organisations reflected on the web.

6.2.1 Live Search's operators and accessibility

The primary reason for the selection of Live Search for the final investigation was the extensiveness of its operators. The introduction of *linkfromdomain* in addition to *linkdomain* allowed for the investigation of inlinks, outlinks and reciprocal links; investigations that are beyond the scope of the other major search engines. Whilst Live Search restricts the number of results that may be retrieved from its database in the same way as other search engines, it also has a smaller index, less accurate results, and saw the removal of the *linkdomain* operator during data collection in the final investigation. Nonetheless, it was still used successfully to gather information about the linking between thousands of different web sites with no disruption to those web sites and using the API in the spirit with which it was provided.

The *linkdomain* and *site* operators, or earlier versions thereof, have been successfully used in many previous webometric investigations (e.g., Almind & Ingwersen, 1997; Thelwall, 2002a; Smith, 2003), however, the introduction of the *linkfromdomain* operator is an important step in making the data accessibility closer to that of data collected with a web crawler, opening up the possibility of investigating outlinks and reciprocal-links. Whilst the investigation of outlinks and reciprocal links are important in their own right, they also offer the opportunity of identifying links that would not necessarily be identified through investigating a web site's inlinks due to the restrictions in the number of results that can be retrieved.

In the link analysis of the ABPI organisational members, 21 of the 139 organisations' web sites included more than a thousand inlinks, whilst Microsoft.com also had over a

thousand outlinks and a thousand reciprocal-links. Without using the outlinks it would theoretically be possible for 21 of the core web sites to each link to one another and none of these links to be found in the search engine results. Incorporating outlinks into the investigation the only relationships that could potentially be missing are the 21 from Microsoft to each of the other organisations. This feature is noticeable in the actual results of the investigation with only one of the core web sites being found to link to Microsoft when looking at Microsoft's inlinks, but 20 of the organisations being found to link to Microsoft when looking at the other core web sites' outlinks.

Whilst utilising the *linkfromdomain* operator to retrieve outlinks, as well as inlinks, reduces the number of potentially missed connections between the core web sites, there are still likely to be many missed connections with web sites not in the core group but in the extended network. It is impossible to determine the number of web sites that were included in the index of the Live Search database that were excluded from the results as the search engine estimates the number of web pages matching the query rather than the number of web sites, and without access to the URLs of the web pages they can not be aggregated into web sites. Nonetheless, the Live Search over-estimation of the number of pages that could not be retrieved due to there being more than a thousand pages returned for certain queries are 32,997,450 pages with inlinks to one of the core web sites, 880,247 pages linked to by one of the core web sites and 104,588 pages MSN reciprocal-linked with a core web site. Although many of these inlinks, as well as all the outlinks and MSN reciprocal-links, are due to the high level of interlinking with microsoft.com, there are still almost one million pages with inlinks to the core network that are not retrieved from the search engine results after taking microsoft.com's figures out of the equation. It seems likely that many of these links will be reflections of collaborations with both external organisations and affiliated organisations which will remain unknown.

The restriction on the number of results that can be retrieved from the web means that the use of search engines for the investigation of inter-organisational web links is more suitable for some studies than others. Whilst it is necessary that an organisation has sufficient web presence for it to link to other organisations, and be linked to by other organisations, if a web site is extremely popular then it is increasingly difficult to retrieve all the links in a search engine's index. Whilst some studies have found that there is insufficient linking between specific organisations or web pages for web links to be a useful source of information (e.g., Kretschmer & Aguillo, 2004), it seems more likely that when investigating the linking between a set of web sites and other unspecified web sites, as in this study, that too many links will be a problem. This is not to say, however, that search engines are of no use in investigating the relationships of the more popular web sites, in fact there is a greater chance of the web site appearing in the results of the top thousand inlinks of another web site as the ranking systems of the major search engines are heavily influenced by a web site's link structure. The problem of too many results seems likely to increase as search engines are able to index more of the web than ever before and more people go online.

Within this investigation an ethical approach was taken to the link analysis, respecting the limitations of the search engine rather than looking for ways of mining the information available. Whilst others have chosen to use increasingly sophisticated queries to collect more of the data than was originally intended (e.g., Bar-Ilan & Peritz, 2007; Thelwall, Forthcoming), the recent reaction of MSN in removing the *linkdomain* operator would seem to suggest that it is best to err on the side of caution and respect the spirit of the limitations imposed by the search engines.

The other limitations with using the Live Search API are the smaller corpus of indexed web pages, which is returned to in the discussion of using a distribution of results approach to investigating the sufficiency of a search engine's crawl (see section 6.2.2), and the additional work necessary in finding the original outlinking page when using the *linkfromdomain* operator.

The extra step necessary in finding the original outlinking page, as well as the subsequent removal of the *linkdomain* which made the extra step more difficult, serves as a reminder that whilst search engines can be useful tools for webometric investigations, the webometricians' plight is not the concern of the search engines. The removal of the *linkdomain* command means that, at the time of writing, it is no longer possible to investigate reciprocal links, data about inlinks would have to be gathered from either Yahoo or Google whilst data about outlinks would have to be gathered by Live Search and would not necessarily be identifiable as a link from a specific page because of the different sets of web pages the different search engines have indexed.

6.2.2 Using distribution to investigate the sufficiency of a search engine's crawl

The difficulty in determining the extent to which a search engine has indexed a particular portion of the web is that we only know what has been indexed rather than what is there to be indexed. This study has attempted to overcome this problem by investigating how the distribution of the numbers of web pages, inlinks, outlinks, and reciprocal links indexed differs from an expected Zipfian distribution. The results show that the estimated number of results differs substantially from the actual number of results, although where there are additional deviations from a Zipfian distribution these are usually reflected in an unexpected web presence rather than the search engine having a heavy bias in favour of one web site or against another. The smaller corpus of Live Search data doesn't seem to favour one particular web site over another in this study, with the possible exception of microsoft.com.

The implications of a search engine providing inaccurate estimated numbers of pages found for a particular query depends very much on the purpose for which the data is being used. If the number of results are being used to compare between two sites, and more than a thousand hits are found for each site, the authority with which it can be stated that one query has a greater number of hits than the other depends on the difference between the estimated number of hits and how far the estimates are from the depth of the results; conclusions should only be drawn with great care as the results are not necessarily either linear or reflect rank order. The effect of this discrepancy within the link analysis of the ABPI member web sites is limited as it was the retrieved results that formed the main basis of the investigation rather than the estimated number of results. The discrepancy between the estimated number of results and the actual number of results does, however, affect the trend line of least squares. Without the discrepancy the difference in the slope of the trend line would be much closer to the Adamic and Huberman's (2002) findings of a slope of -1.

Within this investigation it is only necessary for the researcher to be sure that sufficient numbers of pages have been indexed, rather than it being necessary to be aware of the exact number of pages that are indexed by a search engine. Whilst a researcher's perception of what is or is not sufficient will depend heavily on their own knowledge of the web sites under investigation, placing them in rank order allows for conclusions to be drawn based on the results of the other web sites in the set. When plotting the number of links and

pages indexed by Live Search for each of the 139 web sites in the final study in rank order on a log-log graph, ignoring the drop caused by the discrepancy between the estimated number of results and the actual number of results, there are two noticeable anomalies: the higher than expected number of pages indexed for microsoft.com, and the sudden drop in the number of pages indexed for the lowest ranked web sites.

Whilst a Zipfian distribution may be expected for all the sites on the web (Adamic, 2002), or a sufficiently large random sample thereof, the web sites in this investigation are not a random sample, but rather those organisations in the pharmaceutical industry that are organisational members of the ABPI. The sudden drop-off in the number of pages indexed, inlinks and outlinks for the lower ranked web sites would seem to suggest that the original group of organisations is biased in favour of the larger pharmaceutical organisations, or rather, larger organisations that consider themselves to have an important role in the pharmaceutical industry. An increase in the number of small pharmaceutical organisations, the majority of which probably don't find organisational membership of the ABPI cost effective, would prevent the sudden drop in numbers. Investigation of those sites with a small number of web pages indexed finds that for most it is a reflection of choice in web design, either having a Flash-based site or having a small site that has links to additional affiliated web sites with different domain names. No site was found that had a large number of HTML web pages that only had one or two indexed.

In comparison to the lowly ranked web sites, microsoft.com was found to have an exceptionally large number of inlinks, outlinks, reciprocal links, and pages indexed. There are two potential explanations for this, either the microsoft.com web site is more thoroughly indexed by Live Search because it is part of the same parent organisation, or the web presence of organisations within the computing industry is so much higher than the pharmaceutical industry that when one of the top ranked computing web sites is placed with pharmaceutical organisations it is necessarily an anomaly even if it has been indexed to the same extent. Whilst both are possibilities, that the web site is one of the top ranked computing web sites is definitely likely to be a contributory factor.

Investigation of the anomalous results, highlighted by the plotting of the rank ordered results on a log-log graph, finds reasonable explanations for the differences in the distributions of the results from a Zipfian distribution. Whilst this in no way concludes that the relevant web space has been exhaustively indexed, it suggests an evenness of indexing across the relevant web space. Rather than proving the sufficiency of a crawl, the distribution of results can potentially highlight areas requiring further investigation.

6.2.3 The lack of an alternative data collection source

The problems with using commercial search engines in webometric research have been well documented for a number of years; however, there still seems little indication of an alternative data collection method being made openly available for researchers in the field. Whilst there have been calls for an open academic search engine that would allow webometricians to have full access to the data (e.g., Bar-Ilan, 2001), these search engines have so far failed to appear. Those search engines that have emerged from within the academic community for the use of researchers, such as Google (Snyder & Rosenbaum, 1999), have since moved to the commercial sector and now provide limited access to researchers. Whilst many researchers have utilised web crawlers for webometric investigations, investigating the link relationships between organisations requires access to more link data than it is possible for a small research

group to collect. Within this investigation, whilst the inlinks and outlink data was only collected for 139 web sites, the extended network consisted of a network of 29,289 identified web sites; a number that is liable to be much larger in reality, lowered by Live Search only allowing access to the top 1,000 results. In addition, without the use of a search engine it would be impossible to identify all the web sites that have inlinks to a particular web site. Even if it were possible for a group of researchers to identify and crawl such a large collection of web sites, it would be difficult to ethically justify.

Whilst only looking at the linking between a predetermined set of organisations has been found to leave many web sites unconnected with the other organisations in the predetermined set, both in the pilot study into the West Midlands automobile industry (see section 3.5) and the principal investigation into the pharmaceutical industry, the principal investigation has also found that many of the highly connected web sites, those that have the highest closeness centrality and the highest number of inlinks and outlinks, are not those organisations that are initially identified at the start of the investigation, but rather are discovered through the use of the search engine. The importance of including these sites depends largely on the exhaustiveness of the original set of web sites.

6.3 Investigating link placement

6.3.1 Web presence of the UK pharmaceutical industry

The final study has found the UK pharmaceutical industry to have a well developed web presence; developed enough, at least, to make it worthy of link analysis. The vast majority of the ABPI organisational members were found to have a recognisable web site, and with a network of 29,289 nodes being created from the original seed set of web sites there are sufficient links and organisations to be worthy of investigation (the extent to which these links can be used to provide a useful source of information about the relationships between these organisations is discussed in section 6.3.2). That is not to say, however, that the investigation of the pharmaceutical industry was not without its problems, primarily caused by the identification and operationalization of the concept of a web site.

As well as it being important to identify the key organisations in a particular investigation, it is equally important to identify the numerous domain names that are utilised by the different organisations. Whereas previous link analysis investigations have primarily operationalized the notion of a web site lexically based on one or two domain names per organisation, this seems inappropriate outside the academic community, where there are not necessary one or two predominant web sites, but rather an organisations' web presence may consist of a multitude of small web sites. Within the UK pharmaceutical industry it is not unknown for each product in an organisation's portfolio to have a different web site. This is reflected in the number of MSN reciprocal-links that reflect a collaborative relationship with an affiliated web site.

The decision to operationalize the concept of a web site lexically at the domain level has resulted in the inclusion of certain web sites amongst those with the highest closeness centrality, and those that link to the highest proportion of core web sites: blogspot.com, live.com, and geocities.com. Rather than individual web sites they are blogs and homepages hosted by blogspot.com, live.com, or geocities.com. Whilst collectively these sites may play an important collaborative role, it is unlikely that any single site within these collections has a pivotal role in the pharmaceutical industry.

There is no simple solution to the problem of identifying and operationalizing the concept of the web site, instead it is necessary to deal with web sites on a case by case basis. Even then it may not be possible to ascribe a single rule across a whole web site. For example, whilst certain sections of the live.com domain plays host to numerous independent web sites, other extensive sections can be considered the official live.com site.

6.3.2 Links can reflect collaboration

The classification of a random sample of the web links shows a higher proportion of links to reflect collaborative relationships than has been found in previous investigations, with 66.1 % of outlinks found to reflect a collaborative relationship between the linked web sites, whereas previous investigations of university web sites had found those outlinks reflecting a collaborative relationship to be in the minority (see sections 3.3 and 3.4). In addition, the proportion of inlinks found to reflect a collaborative relationship was also relatively high at 40.8%, and the proportion of MSN reciprocal-links reflecting a collaborative relationship reiterated the findings of the pilot study into the MSN reciprocal-links (see section 3.6) with 92.4% being found to reflect a type of collaborative relationship, in comparison to 93% in the previous study. It is important to note, however, that conclusions about the proportion of all the links that reflect collaboration can not be simply extrapolated out from this sample as there seems to be a wide diversity in the proportion of links that are placed for different reasons across different web sites, with some of the most highly connected web sites having all their links reflecting non-collaborative reasons.

Within this investigation, as in classification exercises in the preliminary investigations (see sections 3.3 & 3.4) and previous investigations (Wilkinson, Harries et al., 2003), the classification was based on a sample of ten web links from each of the web sites to prevent bias in favour of the larger web sites. However, classification of a selection of links from those sites that are most heavily connected with the network shows that the vast majority of them do not reflect collaborative relationships between the linked web sites. So, in effect, limiting the number of links to or from a particular web site may be artificially inflating the proportion of links that reflect a collaborative relationship; creating an artificial construct that exaggerates Van Raan's criticism of citing behaviour analysis, where the analysis of citation behaviour deals primarily with the modal papers whilst citation analysis of research performance deals with the highly cited papers (Moed, 2005). The effect of the heavily connected web sites' tendency to have more links reflecting non-collaborative reasons is unclear from this investigation as the classification sampling technique restricted the number of links from any single core web site, ignoring the web site that the core site was linking with, and those core web sites that were found to be highly connected were found to have links reflecting a range of both collaborative and non-collaborative relationships.

Whilst the classification results highlight the potential of the web as a source of information about organisational collaboration, at least within the UK's pharmaceutical industry, it also reiterates the need for a classification of web links, thus limiting the size of any potential investigation. It seems appropriate for such a classification to be extensive for those sites that are found to have a high proportion of collaborative links, and where there are few collaborative links found within an initial sample, extensive classification should not be carried out.

There are no simple solutions to determining whether or not a web link between two organisations reflects a collaborative relationship between those organisations without

investigating the content of the web pages. Whilst limiting the selection of those web sites involved in the study to a predefined list of likely collaborative organisations may reduce the chances that the link reflects a non-collaborative relationship, and a more extensive collection of an organisation's different URLs is likely to reduce the number of links that are in effect self-links and increase the chances of the web sites having links between them, it is still necessary to investigate the reason for link placement as the primary purpose of link placement is not collaboration.

6.3.3 Hyperlink Network Analysis of the pharmaceutical web space

The centrality of each of the nodes in the extended network was calculated using standard hyperlink network analysis techniques. This allowed for a classification of the links to those web sites that had a high centrality so that it could be investigated whether the reason for the high level of centrality was because they were found to have either a high level of collaboration, or whether most of the links were for non-collaborative reasons. The different types of centrality place different emphases on different aspects of a node's position in the overall network: degree centrality emphasises the importance of the immediate connections with those around them, betweenness centrality if it lies on the geodesics between other nodes, and closeness centrality based on a node's ability to interact quickly with all other nodes (Wasserman & Faust, 1994). Whereas all the nodes with the highest levels of degree centrality were from amongst the original core set of 139 web sites and only one of the nodes with one of the highest levels of betweenness centrality was from the extended network, only one of the top ten nodes with the highest degree of centrality was from the core network.

That the core web sites had the highest degrees of centrality was expected as they could theoretically be found to be linked to a thousand different sites via inlinks, an additional thousand different sites by outlinks, and a yet another thousand different web sites by MSN reciprocal-links. Whereas the additional web sites found in the extended network could only ever be found to link to the original core 139 web sites. With the core networks being directly connected to more of the nodes than those sites in the extended network, it is unsurprising to find that they also dominate the top ten highest betweenness centralities, the inclusion of the .abpi.org.uk web site may be seen as the result of the selection of the original set of core web sites from their web site, and it would have been more appropriate to include their web site as one of the core web sites. The first opportunity for the additional nodes identified in the extended network to have one of the highest levels of centrality is in calculating the closeness centrality, at which point they come to dominate the top of the list.

The classification of a sample of web links for those web sites with the highest levels of centrality necessarily focuses on the direct links with the highly central web sites rather than analysing the additional links in the network that also play an important role in a web site's betweenness centrality or closeness centrality. The classification finds that whilst the majority of the links of the highly central core web sites reflect a mixture of collaborative and non-collaborative relationships, those web sites with a high closeness centrality identified in the extended network have links predominantly reflecting non-collaborative relationships. This raises the question of whether there is any advantage in investigating the relationships between the original core set of web sites and the vast majority of those web sites outside the core set, the answer to which depends primarily on how inclusive the core set of web sites are of those organisations that are liable to have collaborative relationships with the key organisations that form the basis of an investigation. Within this investigation the focus has

been on the relationships in the UK's pharmaceutical industry, a large multinational industry where the 139 core web sites may be thought to be a small selection of the possible web sites that could be included. As such, it may be considered necessary to also investigate the relationships with additional web sites identified in the extended network.

That more outlinks were found to reflect a collaborative relationship than inlinks is not surprising for a set of organisations that were primarily commercial, whilst that a higher proportion of MSN reciprocal-links reflected a collaborative relationship than both inlinks and outlinks was also to be expected due to its restrictive nature. The increase in the proportion of links that reflect a collaborative relationship is principally due to an increase in the proportion of collaborative relationships that are with an affiliated organisation and is at the expense of the proportion of collaborative relationships with non-affiliated web sites, as well as at the expense of the number of non-collaborative relationships reflected.

Whilst identifying the web sites of all the important organisations in the investigation of a whole industry may be an impossible task, it is more feasible for a more narrowly focused investigation, such as the UK West Midlands automobile industry. Whilst the pilot investigation into the UK West Midlands automobile industry (see section 3.5) found a lack of interlinking between some of the web sites, the introduction of the *linkfromdomain* operator allows for the investigation of the relationships between more web sites than was previously possible, and greater attention to the selection of web sites at the start of the investigation is likely to reduce the number of key organisations not included within the study.

6.3.4 Information is the primary purpose of link placement

Both the classification of the reasons for link placement and the application of the SNA measure of closeness centrality emphasise the role of the web link for the highlighting of information: the majority of links reflect non-collaborative relationships, and the organisations with the highest closeness centrality are primarily those that are involved in the dissemination of information. Cozzens (1989) said that we should think of citations "first as rhetoric and second as reward", however when discussing web links we can only say that they should be thought of first as the broader term 'information' with no reason attributable, although sometimes the reason is collaboration. Links are primarily placed to highlight more information, whether this is information on other pages of the same web site, or on alternative web sites. Whilst the reason for highlighting this information varies according to the web site the link is placed on, the vast range of different web sites and purposes that the web is used for means that we can not be more precise about the dominant types of reasons for highlighting information. We do know however that sometimes these links represent collaboration between the linked web sites, and through analysing these web links we can find information about the relationships between organisations that form the basis of the knowledge-based innovation systems that are thought to play such an important role to the modern economy.

6.4 A new source of new information about organisational relationships

One of the most important findings of this investigation is that not only are web links a new source of information about the collaborative relationships between organisations, but that they are a source of new information. The collaborative relationships that web links reflect are, for the most part, relationships that are not visible when searching for multiple organisations in the assignee name field of the patents applied for in the United States Patent

and Trademarks Office (USPTO), or the organisational name field of the scientific articles indexed in the Web of Science (WoS). Equally, few of those collaborative relationships identified in the assignee name field of the USPTO or the organisational name of the WoS were found to be represented by links between the organisation's web sites. There are a number of reasons why the different sources may be showing different organisational relationships: the nature of the collaborative relationships is different; the time frames of the different sources are different; the collaborative relationships are shown, but are not recognised.

The collaborative relationships that result in more than one organisational name in a scientific article, or in more than one assignee name in a patent, are likely to be much more formal than the relationships from web links that were classified as reflecting a collaborative relationship in this investigation. Whilst operationalizing collaboration though the appearance of multiple organisational names within specific bibliographic fields produces more clear cut distinctions as to whether two organisations are collaborating, on the web these clear cut distinctions are not so apparent. Whilst some links may be clearly collaborative with the two linked organisations stating that they are working together on a particular project, others are less clear cut; the use of commercially available software resulting in a web link is unlikely to be judged to be reflecting a collaborative relationship, but without prior knowledge of the software it is not necessarily a simple task to determine whether it is an off-the-shelf solution, or a specially designed piece of software for a specific task. Rather than 'collaborative relationship' and 'non-collaborative relationship' being clearly distinct categories web links may be viewed on a sliding scale between these two points. Although this causes difficulties in achieving high levels of inter-classifier consistency, it also allows greater flexibility in drawing conclusions about what sort of collaborative relationship is being reflected.

Part of the reason for the lack of crossover between those collaborative relationships reflected within the WoS and the USPTO is also likely to be attributable to the time frames that are being examined and the changing nature of the web in comparison to the relative permanence of bibliographic databases. It is possible that the relationships that are found within the bibliographic databases were once visible on the web, but have since disappeared, whilst the relationships currently visible on the web will appear within the bibliographic databases in the future, such longitudinal investigations are a potential avenue of future research.

It is also possible that there are many additional collaborative relationships that are reflected in bibliographic databases and on the web that were not recognised due to this investigation necessarily focusing on the collaborative relationships exhibited within the organisational name field and the assignee name field. Whilst it seems highly likely that citations often reflect real world relationships, there is little way of discerning which citations do and which don't. Web sites, in contrast, do not have to adhere to any particular content standards and are therefore able to provide additional details about the linked organisations.

The suitability of the different sources of information depends on the purpose of the investigation, and how comprehensive it needs to be. Whilst the most comprehensive investigations should use the information available in patent databases, bibliographic databases, and on the web, such extensive investigations require a lot of work to bring the disparate formats together, with the problems of variation in assignee name and organisational name in traditional databases being equally as difficult as grouping together the URLs that belong to an organisation. Part of this problem stems from the idea of organisations having

distinct borders, in reality the boundaries are blurred, and this is reflected in the problems of identifying organisational names, assignee names, and URLs.

6.5 Investigating other sectors

Whilst the final study found the UK pharmaceutical industry to have a well developed enough web presence to make it worthy of link analysis, and a significant enough proportion of the web links were found to reflect collaborative relationships, the UK pharmaceutical industry was chosen because its web space was likely to be developed and to reflect collaborative relationships. Different disciplines in different countries are liable to have different sorts of web presence, and the suitability of the web as a source of information about organisational collaboration, and a search engine as a source of data collection, will vary. Whilst it is known that different disciplines use the web in different ways (Kling & McKim, 2000), further investigation into the reasons for link placement in different disciplines would be necessary before conclusions could be drawn about the use of the web as a source of information about organisational collaboration; greater use of the web does not necessarily equate with greater reflection of organisational collaboration.

6.6 Web links as a new source of information about knowledge-based innovation systems: A microscopic link analysis case study approach

The principal investigation into the web manifestations of knowledge-based innovation systems has found that web links can provide a new source of information about the complex relationships that exist between organisations in the modern economy, although, as with the preliminary investigations, it has shown that there are no simple solutions for either collecting the data or interpreting the data. There are likely to be many links that exist between web sites that are not identified, and those links that are identified need to be examined individually if conclusions are to be drawn about the nature of the relationships between organisations from the web sites with any confidence. It is therefore only suitable for the investigations to be carried out on a small scale, and any conclusions to be couched with numerous caveats.

Within this section a framework is provided for a microscopic link analysis approach to investigating the relationships between organisations. Whilst it is broadly similar to the link analysis methodology proposed by Thelwall (2004a), the limitations in terms of what such an investigation may be expected to show are reflected in the need for an extensive classification and the lack of validation through correlations with an external indicator of innovation.

- A suitable case study - Web links are only a suitable method for investigating the relationships between organisations when looking at a small enough number of organisations so that the web links can be examined individually. Whilst this necessarily excludes the investigation of many multinational organisations, or the investigation of whole industries, there is still a role for its use in small case studies of a specific field in a restricted geographic region. The increase in globalisation has gone hand in hand with a growing recognition of the importance of regional innovation systems (Cooke et al., 2002), and a microscopic link analysis case study provides one method of investigating the nature of the relationships between these organisations. Where a larger investigation is required, classification of a proportion of web links may

indicate which sites are likely to have a significant proportion of links representing collaboration, but the amount of content analysis would soon make the methodology inappropriate.

- Identification of web presence - One of the key issues in the data collection process for the successful use of web links to provide information about the linking between different organisations is the appropriate operationalization of the concept of the web presence; successfully identifying all the different domains that combine to create an organisation's web profile. The final study has shown that it is necessary to look beyond the one or two domains that have previously been thought sufficient for investigating an academic web site's presence. Due to the difficulties in identifying all of the URLs at the start of an investigation the domains that make up a web site's presence should not be thought of as fixed, but updated throughout the investigation as additional domains names are recognised during the classification process.
- Data collection - One of the consistent themes of this thesis has been the importance of an ethical approach to link analysis investigations, an aspect that has been downplayed in previous investigations where the researchers have taken a more cavalier approach in using whichever methodology provides access to the most suitable data and that they can get away with. Within an ethical investigation it is important to minimise the risk to the web site owners and to use the web tools that are made available by the search engines in the manner that they were intended. Selection of the most appropriate data collection tool depends largely upon those tools available at the time of the investigation. Whilst the introduction of the *linkfromdomain* operator by Live Search seemed to favour the use of the search engine for most investigations, the subsequent removal of the *linkdomain* operator has left the most appropriate tool far less obvious. Factors that are likely to favour one of the data collection methods over the other include: the size of the web sites involved in the investigation; search engine coverage of the web sites that are to be included in the investigation; and how assured the researcher is of identifying the most important organisations before collecting the link data. It currently seems appropriate to use Live Search for collecting outlinks, and Yahoo's search engine for collecting inlinks.
- Data cleaning - Whilst it is necessary to aggregate web links, a global approach is not necessarily best; rather, the aggregation needs to be sensitive to the difference between individual web sites. Whilst this is impossible with large scale link analysis, which has necessarily applied a single aggregation method across the board before identifying anomalies, a microscopic approach allows for the aggregation to be based on a site by site basis.
- Determining the meaning of the web links - Classification of web links is the only possible way of determining with any level of confidence whether a link may be taken

as a reflection of collaboration between organisations or not. As there is usually a lack of an identifiable author for most web pages, such classification should use content analysis techniques.

The methodological application of the framework outlined above to small case studies is liable to provide useful information about the collaborations between different organisations that would not be identifiable from traditional sources. Such studies seem most appropriate in the gathering of competitive intelligence on single organisations, rather than drawing conclusions about the state of a whole industry or country, due to the time taken to analyse large quantities of data.

7 Conclusions

7.1 Introduction

The initial aim of this thesis was to determine whether link analysis of the web can provide a new source of information about knowledge-based innovation systems in the UK: developing an appropriate data collection methodology; determining what can be inferred about inter-organisational relationships from web links; and exploring the extent that the information found is new. This chapter summarises the original contributions of the research, discusses the extent to which this investigation has met the original objectives, discusses the potential of the web links as manifestations of knowledge-based innovation systems, and proposes potential avenues for future investigations building on the work of this thesis.

7.2 Original research contributions

This investigation into the web manifestations of knowledge-based innovation systems in the U.K. has made a number of original research contributions. Most importantly, after raising the issue of ethics in web crawling (see section 3.2), it has shown that data may be collected ethically to provide information about the interconnections between web sites of various different sizes and from within different sectors of society (see section 6.2), and that the web provides new information about the relationships between organisations, rather than just a repetition of the same information from an alternative source (see section 6.4). Applying ethical frameworks to an area of study that has previously taken a more cavalier attitude may be seen as an essential step in the legitimizing of a discipline, whilst comparing the results of a link analysis investigation with those of a traditional bibliometric investigation shows the true potential of link analysis.

In addition, investigations documented within this thesis have applied the existing link analysis methodologies of classification and correlation to new areas of the web: investigating the reasons for linking between academia and other sections of the web (see sections 3.3 & 3.4); and investigating the reasons for linking between web sites from different sections of society (see sections 3.5 & 3.6). These studies show that there are significant differences in the linking practices of web sites within different sectors at the microscopic level. The study also investigated an area that has not been looked at before, reciprocal links, showing that they provide a better indication of collaboration than uni-directional web links, although primarily with affiliated sections of the same organisation (see sections 3.6).

7.3 Meeting the objectives of the original investigation

7.3.1 Determining an appropriate data collection methodology

Collection of web data is difficult due to the size (Guillo & Signorini, 2005), the constantly changing nature, and the structure (Broder et al., 2000) of the web, and whilst data has been successfully collected for previous webometric investigations into the academic community, collecting data about the large number of organisations from different sectors that contribute to knowledge-based innovation systems has its own additional difficulties. However, despite these difficulties, it is possible. This study puts forward a successful data collection methodology utilising the latest operators and accessibility offered by the major search engines.

Investigating the actual links between web sites, as opposed to the numbers of links to and from web sites, has previously been dominated by the use of web crawlers. Whilst such methods may be appropriate for investigations into interlinking within the academic community where there are a relatively small number of easily identifiable large web sites with sufficient bandwidth for a single crawler to not cause disruption, the initial preliminary investigation into the ethics of web crawling (see section 3.2) demonstrated that such crawling is not suitable for all situations. Web crawlers can cause denial of service to other users and additional costs to the web site owners unless used with extreme care. Therefore crawlers should only be used when there is no satisfactory alternative, and then sites shouldn't be crawled too quickly, not visited too often, and the crawl should be monitored so that the crawler doesn't get stuck in 'spider traps', large collections of duplicate or almost meaningless web data. These restrictions, unsurprisingly, cause difficulties when investigating the linking between organisations from the different sectors and of different sizes that contribute to knowledge-based innovation systems.

In addition, web crawlers require the complete set of web sites of interest to be identified at the start of the investigation. Whilst this is relatively simple when investigating academic interlinking, it is far more difficult when investigating knowledge-based innovation systems when the purpose of the investigation is identifying the key players that may or may not be expected players in a field. This is clearly seen in the first pilot study into the UK West Midlands automobile industry using URL citations (see section 3.5), although the organisations were all in a small geographic region, and the automobile industry was considered important to the region there was relatively little interlinking between the organisations, with none between those from the industry sector. The lack of links is likely to be a combination of both the different linking cultures for different types of organisation, with the industry sector having few outlinks (Shaw, 2001), and not all of the relevant organisations being included within the investigation. The second pilot study, an investigation into MSN reciprocal links confirmed that there were indeed many outlinks from the web sites which were to web sites not included within the original data set. The only way to use web crawlers to gather sufficient data about the interlinking between organisations which are not necessarily all known at the outset, is to crawl the whole of the web, or the sub-section thereof, e.g., all the pages within a single country-code top-level domain. However, such a crawl would be beyond the capabilities of most webometric research groups, and once again ethically undesirable.

The introduction of application program interfaces (APIs) by the major search engines, i.e., Google, Yahoo, and Live, has enabled the utilisation of a search engine's web data in an ethical manner, as opposed to the scraping of the HTML pages that has been used in other investigations (e.g., Larson, 1996; Heimeriks et al., 2003). Search engine data has been successfully used within the two pilot studies and the principal study to provide information about the interlinking between web sites, initially with a finite set of known web sites using the Google API and operators (see section 3.5), and then in the last two studies investigating the interlinking between a known set of organisations with unknown organisations using the Live Search API and its more extensive set of operators (see sections 3.6 and 4.3.2).

The introduction of the *linkfromdomain* operator by Live Search was initially greeted as a turning point in webometric data collection, bringing the manipulability of search engine data closer to that of data collected with web crawlers. Allowing the identification of inlinks, outlinks and reciprocal-links for the first time as well as limiting the effect of search engines only allowing the first thousand results to be retrieved. However, the subsequent removal of the *linkdomain* operator has limited Live Search's use, and future investigations will require

the combining of the results from more than one search engine; not necessarily to increase coverage, but to be able to view both outlinks and inlinks. The investigation of so-called MSN reciprocal links are, for the foreseeable future, consigned to the pages of this thesis.

Whilst certain of the criticisms of using search engines for webometric investigations remain, such as the lack of consistent results, consistent operators, open crawling and ranking policy, search engines have been shown to provide useful information where there is no acceptable alternative. When investigating information about the web it is necessary to accept a lack of certainties as it is impossible to get a true impression of the whole web, however, it is important that we continue to use additional methods to identify the limitations of our tools. In the final study into the linking between organisational members of the Association of the British Pharmaceutical Industry it is shown that the distribution of the results from a search engine can be used to investigate a search engine's coverage in addition to the actual number of results. Whilst Live Search was found to have sufficiently crawled the large pharmaceutical organisations in the UK, when investigating smaller organisations, across different countries, it becomes increasingly necessary to investigate a search engine's coverage as there is less chance of even coverage.

7.3.2 Determining what web links represent

For the web to provide useful information about inter-organisational relationships it is not only necessary for there to be links between the organisational web sites but for it to be shown that such links are a reflection of organisational relationships. The suggested validation methods for link analysis are a combination of a classification of a sample of web links and an investigation of whether there is a correlation between link counts and an external measure of the inference that is being made (Thelwall, 2004a). However, it is shown within this thesis that because of the differences in the way the web is being used by different organisations within the different sectors that contribute to knowledge-based innovation systems, rather than being able to create an indicator of an organisation's triple-helix-ness for correlation studies, as Boudourides et al. (1999) attempted, classification of a specific link is the only way of determining whether the link represents collaboration, and it is generally for the highlighting of additional information rather than explicitly referring to a collaboration.

Throughout the preliminary investigations, as well as the principal investigation, classification exercises have shown that when investigating web links at the micro level the majority are found to be placed to highlight information rather than reflect collaboration. The classification of the reasons for links to government web sites placed on university web sites found that of the 382 links that were successfully classified only 42 reflected a collaborative research relationship, just 11% (see section 3.3), whilst the more extensive classification into the reasons for university link placement to many different top-level domains equally found the vast majority of links to reflect non-collaborative relationships (see section 3.4), although varying from nine to 27 percent depending on the domain under investigation. Although the final investigation found that for the selected group of pharmaceutical organisations 66.1% of outlinks reflected a collaborative relationship, the vast majority of the links were inlinks, and continued to reflect a non-collaborative relationship (see section 5.4). Whilst it is not necessary for all, or even a majority, of links to represent collaboration, for web links to provide a useful indicator of an organisation's triple-helix-ness, after determining correlation is not a suitable method of validation, the classification process takes on additional importance.

Whilst the university-to-government preliminary study (see section 3.2) included an investigation into whether or not there was a correlation between an institution's outlinks per member of staff and research productivity, the outlinking institutions may be considered to have sufficiently similar linking behaviour for variations in the results to reflect a real world difference, whilst the RAE gave an external indicator with which to compare against. In comparison, there is not an external indicator of triple-helix-ness which can be equally applied across the different sectors to compare against, and variations in linking behaviour is liable to be a result of differences in the linking behaviour of different types of organisation with different priorities.

Despite classification exercises showing that web links do not generally reflect collaboration, at least explicitly, patterns of real world relationships have been seen to emerge in the linking between the different levels of local government when using URL citations (see section 3.5). However, these relationships were given the opportunity to emerge due to the finite number of web sites that formed the basis of the investigation. When allowing the unfettered inclusion of unevaluated web sites, as with the web sites in the principal investigation that were not selected but rather were found to link to the core web sites (see section 4), the few collaborative links are quickly overshadowed by the mass of links based on information purposes.

Whilst conclusions can not be drawn about the reasons for link placement, or the dominance of a web site within a network, without investigating the site or the links more closely, the informal nature of the web means that web pages often highlight collaborations of various levels of formality in a way that wouldn't be so obvious in traditional bibliometric sources.

7.3.3 Determining the difference between webometric data and traditional bibliometric data

The final investigation has successfully shown that there are distinct differences between the information about inter-organisational relationships available on the web, and the information about inter-organisational relationships contained within the organisational field of the Web of Science's bibliographic records, and the assignee field of the US patent office: collaborations that are shown in the links between web sites that are not listed in traditional databases; and collaborations in traditional databases that are not reflected in web links between the collaborating organisations. Whilst the results are limited to the UK pharmaceutical industry, it would seem highly likely that the conclusions apply equally to other fields. There are, however, additional connections between patents and science articles which were not investigated within the study of the UK pharmaceutical industry, such as the citations and acknowledgements.

It quickly becomes clear during the classification of web links that they offer the potential to show very different types of collaboration to that explicit within bibliographic databases. Whilst it is difficult to get inter-classifier agreement on the exact nature of a collaborative relationship, when investigating the reason for link placement on university web pages to different domains (see section 3.4) both classifiers found web links that reflected much more informal collaborations, memberships of organisations, one party carrying out work for another, as well as the traditional research collaborations and funding which are more likely to appear in the traditional databases. Although some web sites from the commercial sector are more circumspect in the placing of a large number of web links, the findings of the

final study show that the commercial sector also has a number of collaborations reflected in web links that aren't reflect in the WoS or the USPTO. Whilst it may be that there is more interlinking on pharmaceutical web sites than on the web sites of other commercial organisations, there is also likely to be more patents filed to protect their property rights.

The conclusion that web links can provide information about inter-organisational collaboration that is not available in the traditional bibliometric databases is based on investigating the organisational field name of the WoS and the assignee field name of the USPTO. Whilst it is possible that these relationships are also reflected within citations of patents and journal articles, further labour-intensive research would be necessary, with all the papers cited having to be looked at individually to determine the organisational affiliation of the author. This is not practical for large pharmaceutical organisations that may have thousands of papers and patents, each potentially with tens of citations. Even if a link does reflect information available within a citation it is extremely unlikely that it will be stated in the paper that the citation has been placed in preference to another because the author has worked with the cited author, and is thus more aware of their work and more inclined to cite it. The informal nature of many web sites give the author freedom to link to whomever they like, and provide a description of the relationship to the owner of the linked site.

That inter-organisational relationships are discernable within traditional databases that are not reflected within web links between organisations is more surprising, but by no means astonishing. Rather than the information not being available on the web, it seems more likely that a lot of the information is there, it is just not available in web links. It is not unusual for individuals to list their publications on a web page, but it would be unusual for each of the collaborating authors' names to be hyperlinked to their web page.

Whilst this study has shown that not only is the web a new source of information about inter-organisational collaboration, but also a source of new information, there is a lot more research necessary.

7.4 The potential of web links as manifestations of knowledge-based innovation systems

The thesis has shown that it is possible to collect information about the web links to and from various different organisations within different sectors of society, and with extensive classifications to separate the wheat from the chaff, some of the links can be found to provide a new source of information about inter-organisational collaboration, the types of relationships that have been described as so widespread within innovative companies as to seem essential to the innovation process (Smith, 2005). However, the ability of these links to provide indicators of the key organisations within an innovation system, or even the major flows of information is fairly limited. Not only because of the extensive classification necessary, but also because of the necessarily incomplete picture of inter-organisational collaborations that web links provide, heavily dependent as they are on the different norms in linking behaviour in different sectors and in different disciplines.

With the technology and tools currently available, and our current level of understanding of linking practices, it seems that the most appropriate application of a link-analysis of a knowledge-based innovation system is as a basis for competitive intelligence. Within this field the extensive classification that is necessary to provide useful information about an organisation's collaborations and their position within a system of innovation, it would also be an opportunity to analyse people's opinions of an organisation, and the work

that is generating interest. Whilst hyperlink network analysis provides a method of determining an organisations role within a network, the investigation of the ABPI organisational members shows the necessity of classifying the links so that only those that relate to collaboration are used to calculate the most central web site: in much the same way as Smith (2003) proposed substantive web impact factors, it is suggested here that substantive hyperlink network analysis help inform the role of an organisation within a network.

7.5 Future research

The web is an exciting new source of information about inter-organisational relationships, with many potential avenues of research that can be followed up. However, whilst on the one hand the findings of this thesis may be considered fairly basic, before the field of webometrics can move forward it may be sensible for it to take a step back; to investigate the web at an even more basic level. This should start with a reassessment of how we ascribe traditional notions of organisation, and organisational type, on the web:

- Determining organisational type: Within this investigation organisations have been distinguished by their traditional sector categorisations: academic, industry, or government. However, as the different models of knowledge-based innovation systems acknowledge, such categories are an artificial construct, and in reality we live in a world of hybrid organisations of one degree or another. Rather than investigating the web presence of organisations from different sectors, a more productive approach may be to categorise organisations according to their web presence, including their use of numerous different domain names.
- Theory of linking: Constructing new notions of what types of organisation are on the web provides a new structure for building our understanding of link theory.

Although there is no doubt that useful information can be gathered without improvements to the basics, more work on the basics is necessary if webometrics is to live up to its potential.

8 Bibliography

ACM. (1992). *ACM code of ethics and professional conduct*. Retrieved March 1, 2005, from http://www.acm.org/constitution/code.html

Adamic, L., & Huberman, B.A. (2001). The web's hidden order. *Communications of the ACM, 44*(9), 55-59.

Adamic, L. A., & Huberman, B.A. (2002). Zipf's law and the Internet. *Glottometrics, 3*, 143-150.

Aguillo, I. (1997). *Cybermetrics'97 (Jerusalem, Israel)*. Retrieved May 12, 2007, from http://www.cindoc.csic.es/cybermetrics/cybermetrics97.html

Ajiferuke, I., & Wolfram, D. (2004). Modelling the characteristics of web page outlinks. *Scientometrics, 59*(1), 43-62.

Almind, T.C., & Ingwersen, P. (1996). *Informetric analysis on the World Wide Web: A methodological approach to "internetometrics"*. Centre for Informetric Studies, Royal School of Library and Information Science. (CIS Report 2).

Almind, T.C., & Ingwersen, P. (1997). Informetric analysis 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.

Arrington, R. L. (1998). *Western ethics: An historical introduction*. Oxford: Blackwell.

Arroyo, N. (2004). What is the Invisible Web? A Crawler Perspective. *Proceedings of the AoIR-ASIST 2004 Workshop on Web Science Research Methods, September 19, 2004, Brighton, UK*. Retrieved May 12, 2007, from http://cybermetrics.wlv.ac.uk/AoIRASIST/arroyo.html

AskJeeves. (2005). *Technology & Features: Teoma Search Technology*. AskJeeves inc. Retrieved February 23, 2005, from http://sp.ask.com/docs/about/tech_teoma.html

Baeza-Yates, R., & Castillo, C. (2002). Balancing volume, quality and freshness in web crawling. In A. Abraham, J. Ruiz-del-solar, & M. Köppen (Eds.), *Soft-computing systems: Design, management and applications, frontiers in artificial intelligence and applications* (pp.565-572). Santiago, Chile: IOS press.

Baeza-Yates, R., & Castillo, C. (2004). Crawling the infinite web: Five levels are enough. In S. Leonardi (Ed.), *Algorithms and Models for the Web-Graph: Third International Workshop* (pp. 156-167). Heidelberg: Springer.

Bar-Ilan, J. (1998). On the overlap, the precision and estimated recall of search engines: A case study of the query "Erdos". *Scientometrics, 42*(2), 207-228.

Bar-Ilan, J. (1999). Search engine results over time – A case study on search engine stability. *Cybermetrics, 2/3*(1), Retrieved October 4, 2005, from http://cybermetrics.cindoc.csic.es/pruebas/v2i1p1.htm

Bar-Ilan, J. (2000a). Evaluating the stability of the search tools Hotbot and Snap: A case study. *Online Information Review, 24*(6), 439-449.

Bar-Ilan, J. (2000b). The web as an information source on informetrics? A content analysis. *Journal of the American Society for Information Science, 51*(5), 432-443.

Bar-Ilan, J. (2001). Data collection methods on the web for informetric purposes. *Scientometrics, 50*(1), 7-32.

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

Bar-Ilan, J. (2004a2). Self-linking and self-linked rates of academic institutions on the web. *Scientometrics, 59*(1), 29-41.

Bar-Ilan, J. (2004b). The use of Web search engines in information science research. In B. Cronin (Ed.), *Annual Review of Information Science and Technology*, Vol. 38 (pp.231-288). Medford, NJ: Information Today.

Bar-Ilan, J. (2005a). Expectations versus reality: Search engine features needed for web research at mid 2005. *Cybermetrics, 9*(1). Retrieved May 6, 2007, from http://www.cindoc.csic.es/cybermetrics/articles/v9i1p2.html.

Bar-Ilan, J. (2005b). What do we know about links and linking? A framework for studying links in academic environments. *Information Processing and Management, 41*(4), 973-986.

Bar-Ilan, J., & Echerman, A. (2005). The anthrax scare and the web: A content analysis of web pages linking to resources on anthrax. *Scientometrics, 63*(3), 443-462.

Bar-Ilan, J., & Peritz, B. (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.

Bar-Ilan, J., & Peritz, B. (2007). The lifespan of "informetrics" on the Web: An eight year study (1998-2006). In D. Torres-Salinas, & H. F. Moed (Eds.), *Proceedings of the 11th International Conference of the International Society for Scientometrics and Informetrics*: *Vol. 1*. (pp.52-62). Madrid: CINDOC.

Barjak, F. (2006). The role of the internet in informal scholarly communication. *Journal of the American Society for Information Science and Technology, 57*(10), 1350-1367.

BBC News. (2005). *US duo in first spam*. BBC. Retrieved February 23, 2005, from http://news.bbc.co.uk/1/hi/technology/3981099.stm

Berners-Lee, T., & Fischetti, M. (1999). *Weaving the Web: The past, present and future of the World Wide Web by its inventor*. London: Orion Business Books.

Beaulieu, A., & Simakova, E. (2006). Textured connectivity: An ethnographic approach to understanding the timescape of hyperlinks. *Cybermetrics, 10*(1). Retrieved May 6, 2007, from http://www.cindoc.csic.es/cybermetrics/articles/v10i1p5.html

Björneborn, L. (2004). *Small-world link structures across an academic Web space: A library and information science approach*. Ph.D. Thesis, Royal School of Library and Information Science, Copenhagen, Denmark.

Bordons, M., & Gómez, I. (2003). Collaboration networks in science. In B. Cronin, & H. B. Atkins (Eds.), *The Web of Knowledge: A Festschrift in Honor of Eugene Garfield* (pp. 197-213). New Jersey: ASIS.

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

Borrull, A. L., & Oppenheim, C. (2004). Legal aspects of the web. *Annual Review of Information Science & Technology, 38*(1), 483-548.

Bossy, M. J. (1995). *The last of the litter: "Netometrics"*. Solaris, 2. Retrieved May 6, 2007, from http://biblio-fr.info.unicaen.fr/bnum/jelec/Solaris/d02/2bossy.html

Boudourides, M. A., Sigrist, B., & Alevizos, P. D. (1999). *Webometrics and the Self-Organisation of the European Information Society: Draft report of the SOEIS project*. Retrieved March 12, 2007, from http://www.cindoc.csic.es/cybermetrics/pdf/14.pdf

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.

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

Brooks, T. A. (1985). Private acts and public objects: An investigation of citer motivations. *Journal of the American Society for Information Science, 36*(4), 223-229.

Brooks, T. A. (1986). Evidence of complex citer motivations. *Journal of the American Society for Information Science, 37*(1), 34-36.

Brown, C. (2004). The Mathew Effect of the Annual Reviews series and the flow of scientific communication through the world wide web. *Scientometrics, 60*(1), 25-36.

Buell, K. C. (2000). "Start spreading the news": Why republishing material from "disreputable" news reports must be constitutionally protected. *New York University Law Review, 75*(4), 966-1003.

Butcher, J., & Jeffrey, P. (2005). The use of bibliometric indicators to explore industry-academia collaboration trends over time in the field of membrane use for water treatment. *Technovation, 25*, 1273-1280.

Cabinet Office. (2005). *Transformational government: Enabled by technology*. Retrieved May 10, 2007, from http://www.cio.gov.uk/documents/pdf/transgov/transgov-strategy.pdf

Calluzzo, V. J., & Cante, C. J. (2004). Ethics in information technology and software use. *Journal of Business Ethics, 51*(3), 301-312.

Campanario, J. M. (2003). Citation analysis. In. Feather, J. & Sturges, P. (eds.). *International Encyclopaedia of Information and Library Science 2nd Edition.* Available at: http://www2.uah.es/jmc/ai44.pdf

Carey, P. (2004). *Data protection: A practical guide to UK and EU law*. Oxford: Oxford University Press.

Carlin, A. P. (2003). Disciplinary debates and bases of interdisciplinary studies: The place of research ethics in library and information science. *Library and Information Science Research, 25*(1), 3-18.

Case, D. O., & Higgins, G. M. (2000). How can we investigate citation behaviors? A study of reasons for citing literature in communication. *Journal of the American Society for Information Science, 51*(7), 635-645.

Casey, T. D. (2000). *ISP survival guide: Strategies for managing copyright, spam, cache and privacy regulations.* New York: Wiley.

Cassiman, B., Glenisson, P., & Van Looy, B. (2007). Measuring the industry-science links through inventor-author relations: A profiling methodology. *Scientometrics, 70*(2), 379-391.

Chakrabarti, S. (2003). *Mining the web: Analysis of hypertext and semi structured data*. New York: Morgan Kaufmann.

Chambers. (1991). *Chambers concise dictionary*. Edinburgh: W & R Chambers Ltd.

Chau, M., Shiu, B., Chan, I., & Chen, H. (2007). Redips: Backlink search and analysis on the web for business intelligence analysis. *Journal of the American Society for Information Science and Technology, 58*(3), 351-365.

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., & Rosenthal, M. (1996). Search engines for the World Wide Web: A comparative study and evaluation methodology. *Proceedings of the Annual Meeting of the American Society for Informaiton Science, 33*, 127-135.

Chubin, D. E., & Moitra, S. D. (1975). Content analysis of reference: Adjunct or alternative to citation counting? *Social Studies of Science, 5*(4), 423-441.

Lombardi, C. (2007). *Spam experts at MIT lift curtain on search.* CNET News. Retrieved May 12, 2007, from http://news.com.com/Spam+experts+at+MIT+lift+curtain+on+search/2100-1024_3-6172199.html?tag=alert

Cooke, P., Uranga, M. G., & Etxebarria, G. (1997). Regional innovation systems: Institutional and organisational dimensions. *Regional Policy, 26*(4), 475-491.

CorporateWatch. (2003). *The Association of the British Pharmaceutical Industry*. Retrieved May 1, 2007, from http://www.corporatewatch.org/?lid=397

Costello, P., Garner, S., Homer, G., & Thompson, D. (1999). Can the Internet provide the West Midlands automotive industry with the ultimate lean production tool? - A case study of the Auto lean Project. *The Second International SME Conference*: *Manufacturing and Business Systems Group, 1999, March 29-31, University of Plymouth, UK*.

Cothey, V. (2003). Web-crawling reliability. *Journal of the American Society for Information Science and Technology*, 55(14), 1228-1238.

Cothey, V., Aguillo, I., & Arroyo, N. (2006). Operationalising "Websites": lexically, semantically or topologically? *Cybermetrics, 10*(1). Retrieved May 10, 2007, from http://www.cindoc.csic.es/cybermetrics/articles/v10i1p3.html

Cozzens, S. E. (1989). What do citations count? The rhetoric-first model. *Scientometrics, 15*(5-6), 437-447.

Cronin, B. (1999). The Warholian moment and other proto-indicators of scholarly salience. *Journal of the American Society for Information Science, 50*(10), 953-955.

Cronin, B., & McKim, G. (1996). Science and scholarship on the world wide web: A North American perspective. *Journal of Documentation 52*(2), 163-171.

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

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 Festchrift in Honor of Eugene Garfield* (pp.517-534). Medford, NJ: Information Technology Inc.

Davies, M. (2001). Creating and using multi-million word corpora from web-based newspapers. In R. C. Simpson, & J. M. Swales (Eds.), *Corpus Linguistics in North America* (pp.58-75). Ann Arbor: University of Michigan.

Davis, M. (1991). Thinking like an engineer: The place of a code of ethics in the practice of a profession. *Philosophy and Public Affairs, 20*(2), 150-167.

DirectGov. (2005). *Local councils in the West Midlands*. Retrieved June 12, 2005, from http://www.direct.gov.uk/QuickFind/LocalCouncils/LocalCouncilArticle/fs/en?CONTENT_ID=4003650&chk=BVWOSS

DTI. (2005). *Auto Industry: West Midlands*. Retrieved June 12, 2005, from http://www.autoindustry.co.uk/regions/westmidlands/2

DTI. (2007a). *Innovation*. Retrieved March 12, 2007, from http://www.dti.gov.uk/innovation/index.html

DTI. (2007b). *UK Innovation Survey*. Retrieved May 12, 2007 from http://www.dti.gov.uk/files/file9688.pdf

du Gay, P., Hall, S., Janes, L., Mackay, H., & Negus, K. (1997). *Doing cultural studies: The Story of the Sony Walkman*. London: Sage.

Edgerton, D. (2004). The linear model did not exist. In K. Grandin, N. Worms, & S. Widmalm (Eds.), *The Science-Industry Nexus: History, Policy, Implications* (pp.31-57). Sagamore Beach: Science History Publications.

Education Guardian. (2001). *About the tables.* Retrieved January 5, 2005, from http://education.guardian.co.uk/secondaryschoolsguide/story/0,11228,602663,00.html.

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

Eichmann, D. (1994, October). *Ethical web agents*. Paper presented at the Second International World Wide Web Conference, Chicago, Illinois, USA. Retrieved June 6, 2007, from http://mingo.info-science.uiowa.edu/eichmann/www-f94/ethics/ethics.ps

Ess, C., & Committee, A. E. W. (2002). *Ethical decision-making and Internet research. Recommendations from the aoir ethics working committee*. Retrieved February 23, 2005, from http://aoir.org/reports/ethics.pdf

Etzkowitz, H. (2001). The Entrepreneurial University and the Emergence of Democratic Corporatism. In H. Etzkowitz, & L. Leydesdorff (Eds.), *Universities and the Global Knowledge Economy: A Triple Helix of University-Industry-Government Relations* (pp.141-152). London: Pinter.

Etzkowitz, H., & Leydesdorff, L. (1995). The Triple Helix: University-industry-government Relations. *EASST Review, 14*(1). Retrieved March 12, 2007, from http://www.easst.net/review/march1995/leydesdorff

Etzkowitz, H., & Leydesdorff, L. (2000). The dynamics of innovation: from National Systems and "Mode 2" to a Triple Helix of university-industry-government relations. *Research Policy, 29*(2), 109-123.

Etzkowitz, H., & Leydesdorff, L. (2001). Introduction: Universities in the Global Knowledge Economy. In H. Etzkowitz, & L. Leydesdorff (Eds.), *Universities and the Global*

Knowledge Economy: A Triple Helix of University-Industry-Government Relations (pp.1-8). London: Pinter.

Etzkowitz, H., & Zhou, C. (2006). Triple Helix twins: Innovation and sustainability. *Science and Public Policy, 33*(1), 77-83.

Eysenbach, G., & Köhler, C. (2002). How do consumers search for and appraise health information on the world wide web? Qualitative study using focus groups, usability tests, and in-depth interviews. *BMJ, 324*(9), 573-577.

Fagerberg, J. (2005). Innovation: A guide to the literature. In F. Fagerberg, D. C. Mowery, & R. R. Nelson (Eds.), *The Oxford Handbook of Innovation* (pp.1-26). Oxford: OUP.

Fausett, B. A. (2002). Into the Deep: How deep linking can sink you. *New Architect* (Oct 2002). Retrieved May 12, 2007, from http://www.ddj.com/dept/architect/184411709

Fisher, D. A., & Klein, J. A. (2003). From Mode 1 to Mode 2: Can universities learn from consultancies. *Industry & Higher Education, 17*(1), 45-49.

Fleck, J. (2004). The Structure of Technological Evolutions: Linear models, configurations, and Systems of Development. In K. Grandin, N. Wormbs, & S. Widmalm (Eds.), *The Science-Industry Nexus: History, Policy, Implications* (pp.229-255). Canton, MA: Science History Publications.

Frankel, M. S., & Siang, S. (1999). *Ethical and legal aspects of human subjects research on the internet.* American Association for the Advancement of Science. Retrieved May 12, 2007, from http://www.aaas.org/spp/sfrl/projects/intres/report.pdf

Garfield, E. (1979). *Citation indexing: Its theory and application in science, technology, and humanities*. Philadelphia: ISI Press.

Garfield, E. (1998). Random thoughts on citationology: Its theory and practice. *Scientometrics, 43*(1), 69-76.

Gau, Y., & Vaughan, L. (2005). Web hyperlink profiles of news sites: A comparison of newspapers of USA, Canada, and China. *Aslib Proceedings: New Information Perspectives, 57*(5), 398-411.

Gibbons, M., Limoges, C., Nowotny, H., Schwartzman, S., Scott, P., & Trow, M. (1994). *The new production of knowledge: The dynamics of science and research in contemporary societies*. London: Sage publications.

Glänzel, W., & Meyer, M. (2003). Patents cited in the scientific literature: An exploratory study of 'reverse' citation relations. *Scientometrics, 58*(2), 415-428.

Glänzel, W., & Schubert, A. (2004). Analysing scientific networks through co-authorship. In H. F. Moed, W. Glänzel, & U. Schmoch (Eds.), *Handbook of Quantitative Science and Technology Research* (pp.257-276). Dordrecht: Kluwer Academic Publishers.

Godin, B. (2005). The Linear Model of Innovation: The historical construction of an analytical framework. *Project on the History and Sociology of S&T Statistics. Working paper No. 30.* Retrieved May 7, 2007, from http://www.csiic.ca/PDF/Godin_30.pdf

Google. (2006). *Google SOAP search API (Beta)*. Retrieved May 12, 2007, from http://code.google.com/apis/soapsearch/

Google Librarian Central. (2007). *Back from WebSearch University*. Retrieved May 12, 2007, from http://librariancentral.blogspot.com/2007/05/back-from-websearch-university.html

Gopal, R. D., Sanders, G. L., Bhattacharjee, S., Agrawal, M., & Wagner, S. C. (2004). A behavioral model of digital music piracy. *Journal of Organisational Computing and Electronic Commerce, 14*(2), 89-105.

Gresham, J. L. (1994). From invisible college to cyberspace college: Computer conferencing and the transformation of informal scholarly communication networks. *Interpersonal Computing and Technology, 2*(4), 37-52. Retrieved May 12, 2007, from http:www.helsinki.fi/science/optek/1994/n4/gresham.txt

Grupp, H., & Mogee, M. E. (2004). Indicators for national science and technology policy: Their development, use and possible misuse. In H. F. Moed, W. Glänzel, & U. Schmoch (Eds.), *Handbook of Quantitative Science and Technology Research* (pp.75-94). Dordrecht: Kluwer Academic Publishers.

Gulli, A., & Singorini, A. (2005). The indexable web is more than 11.5 billion pages. *International World Wide Wed Conference* (pp.902-903). Retrieved June 7, 2007, from http:://portal.acm.org/citation.cfm?id=1062745.1062789

Haas, S. W., & Grams, S. E. (1998). A link taxonomy for web pages. In C. Prestion (Ed.), *Proceedings of the 61st Annual Meeting of the American Society for Information Science* (pp.485-495). Medford, NJ: Information Today Inc.

Harries, G., Wilkinson, D., Price, L., Fairclough, R., & Thelwall, M. (2004). Hyperlinks as a data source for science mapping, *Journal of Information Science, 30*(5), 236-447.

Harter, S. P., & Ford, C. E. (2000). Web-base analyses of E-journal impact: Approaches, problems, and issues. *Journal of the American Society for Information Science, 51*(13), 1159-1176.

HEFCE. (2006). *Response to consultation on successor to research assessment exercise.* Retrieved May 12, 2007, from http://www.hefce.ac.uk/News/HEFCE/2006/rae.htm

Heimeriks, G., Horlesberger, 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 hyperlinks networks: The mean of hyperlink based indicators of knowledge production. *Cybermetrics, 10*(1), Retrieved May 12, 2007, from http://www.cindoc.csic.es/cybermetrics/articles/v10i1p1.html

Hine, C. (1998). Virtual ethnography. *IRISS '98 Conference Papers*. Retrieved June 7, 2007, http://www.intute.ac.uk/socialsciences/archive/iriss/papers/paper16.html

Hinze, S., & Schmoch, U. (2004). Opening the black box: Analytical approaches and their impact on the outcome of statistical patent analysis. In H. F. Moed, W. Glänzel, & U. Schmoch (Eds.), *Handbook of Quantitative Science and Technology Research* (pp.215-235). Dordrecht: Kluwer Academic Publishers.

HMSO. (1999). *The Telecommunications (Data Protection and Privacy) Regulations 1999*. Statutory Instrument 1999 No. 2093. Retrieved February 23, 2005, from http://www.legislation.hmso.gov.uk/si/si1999/19992093.htm

Holdsworth, D. (1995). Ethical decision-making in science and technology. In B. Almond (Ed.), *Introducing applied ethics* (pp.130-147). Oxford, UK: Blackwell.

Holton, G. (1978). Can science be measured. In Y. Elkana, J. Lederberg, R. K. Merton, A. Thackray, & H. Zuckerman (Eds.), *Toward a Metric of Science: The advent of science indicators*. (pp.39-68). New York: Wiley- Interscience.

Huizingh, E. K. R. E. (2000). The content and design of web sites an empirical study. *Information and Management, 37*(3), 123-134.

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

Ingwersen, P. (2006, May). *Webometrics: Ten year of expansion*. Invited plenary talk at the International Workshop on Webometrics, Informetrics and Scientometrics & Seventh COLLNET Meeting, Nancy, France. Retrieved May 13, 2007, from http://eprints.rclis.org/archive/00006264/fullmetadata.html

Ingwersen, P., & Björneborn, L. (2004). Methodological issues of webometric studies. In H. F. Moed, W. Glänzel, & U. Schmoch (Eds.), *Handbook of Quantitative Science and Technology Research* (pp.339-369). Dordrecht: Kluwer Academic Publishers.

Internet Archive. (2005). [*Homepage*]. Retrieved January 10, 2005, from http://www.archive.org

Introna, L. D., & Nissenbaum, H. (2000). The Internet as a democratic medium: Why the politics of search engines matters. *The Information Society, 16*(3), 169-185.

Jacso, P. (2005). As we may search- Comparison of major features of Web of Science, Scopus and Google Scholar citation-based and citation-enhanced databases. *Current Science, 89*(9), 1537-1547.

Jankowski, N., & van Selm, M. (2001). *Research ethics in a virtual world: Some guidelines and illustrations*. Retrieved February 23, 2005, from http://oase.uci.kun.nl/~jankow/Jankowski/publications/Research%20Ethics%20in%20a%20Virtual%20World.pdf

Jeffrey, P. (2003). Smoothing the waters: Observations on the Process of Cross-Disciplinary Collaboration. *Social Studies of Science, 33*(4), 539-562.

Johnson, D. G. (2004). Computer Ethics. In H. Nissenbaum, & M. E. Price (Eds.), *Academy & the Internet* (pp.143-167). New York: Peter Lang.

Jones, R. A. (1994). The ethics of research in cyberspace. *Internet Research: Electronic Networking Applications and Policy, 4*(3), 30-35.

Kamada, T., & Kawai, S. (1989). An algorithm for drawing general undirected graph. *Information Processing Letters, 31*(1), 7-15.

Kaplan, N. (1965). The norms of citation behaviour: Prolegomena to the footnote. *American Documentation, 16*(3), 179-184.

Karki, M. M. S., & Krishnan, K. S. (1997). Patent citation analysis: A policy analysis tool. *World Patent Information, 19*(4), 269-272.

Katz, J.S. (2006). Introduction to a special issue on web indicators for innovation systems. *Research Evaluation, 15*(2), 83.

Katz, J. S., & Hicks, D. M. (1996). A systemic view of British science. *Scientometrics, 35*(1), 133-154.

Katz, J. S., & Martin, B. R. (1997). What is research collaboration? *Research Policy, 26*(1), 1-18.

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.

Kline, S. J., & Rosenberg, N. (1986). An overview of innovation. In R. Landau, & N. Rosenberg (Eds.), *The positive sum strategy: Harnessing technology for economic growth,* (pp.275-304). Washington, DC: National Academy Press.

Kling, R., & McKim, G. (2000). Not just a matter of time: Field differences and the shaping of electronic media in supporting scientific communication. *Journal of the American Society for Information Science, 51*(14), 1306-1320.

Koehler, W. (1999). An analysis of web page and web site constancy and permanence. *Journal of the American Society for Information Science, 50*(2), 162-180.

Koehler, W. (2002). Web page change and persistence - A four-year longitudinal study. *Journal of the American Society for Information Science, 53*(2), 162-171.

Koenig, M. E. D. (1983a). A bibliometric analysis of pharmaceutical research. *Research Policy, 12*(1), 15-36.

Koenig, M. E. D. (1983b). Bibliometric indicators versus expert opinion in assessing research performance. *Journal of the American Society for Information Science, 34*(2), 136-145.

Kogan, S. L., & Muller, M. J. (2006). Ethnographic study of collaborative knowledge work. *IBM Systems Journal, 45*(4), 759-771. Retrieved May 12, 2007, from http://www.research.ibm.com/journal/sj/454/kogan.pdf

Koster, M. (1993). *Guidelines for robot writers.* Retrieved February 23, 2005, from http://www.robotstxt.org/wc/guidelines.html

Koster, M. (1994). *A standard for robot exclusion.* Retrieved March 3, 2005, from http://www.robotstxt.org/wc/norobots.html

Koster, M. (1996). *Evaluation of the standard for robots exclusion*. Retrieved February 23, 2005, from http://www.robotstxt.org/wc/eval.html

Kousha, K., & Thelwall, M. (2005). Motivations for URL citations to open access library and information science articles. In P. Ingwersen, & B. Larsen (Eds.), *Proceedings of the 10th International Conference of the International Society for Scientometrics and Informetrics: Vol. 1.* (pp.67-77). Stockholm: Karolinska University Press.

Kretschmer, H., & Aguillo, I. F. (2004). Visibility of collaboration on the web. *Scientometrics, 61*(3), 405-426.

Kretschmer, H., & Kretschmer, T. (2006). A new centrality measure for social network analysis applicable to bibliometric and webometric data. *In* COLLNET (2006) *International Workshop on Webometrics, Informetrics and Scientometrics (10-12 May, 2006,* Nancy. Retrieved January 4th, 2008, from http://eprints.rclis.org/archive/00006351/02/Kretschmer18aps.pdf

Kretschmer, H., Kretschmer, U., & Kretschmer, T. (2007). Reflection of co-authorship networks in the web: Web hyperlinks versus web visibility rates. *Scientometrics, 70*(2), 519-540.

Krippendorff, K. (2004). *Content Analysis: an introduction to its methodology* (2nd ed.). London: Sage.

Krogh, C. (1996). The rights of agents. In M. Wooldridge, J. P. Muller, & M. Tambe (Eds.), *Intelligent Agents II, Agent Theories, Architectures and Languages* (pp.1-16). Marina Del Ray, CA: Springer Verlag.

Larson, R. R. (1996). Bibliometrics of the World Wide Web: An exploratory analysis of the intellectual structure of cyber space. In S. Hardin (Ed.), *Proceedings of the 59th annual meeting, ASIS* (pp. 71-79). Baltimore: Learned Information Ltd.

Lawrence, S., & Giles, C. L. (1998). Searching the world wide web. *Science, 280*(5360), 98-100.

Lazarev, V. S. (1996). On chaos in bibliometric terminology. *Scientometrics, 35*(2), 271-277.

Lazzaroni, M., & Piccaluga, A. (2003). Towards the entrepreneurial university. *Local Economy, 18*(1), 38-48.

Lewandowski, D., Wahlig, H., & Meyer-Bautor, G. (2006). The freshness of web search engines' databases. *Journal of Information Science, 3*(2), 131-148.

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

Leydesdorff, L. (2001). Indicators of innovation in a knowledge-based economy. *Cybermetrics, 5*(1). Retrieved May 12, 2007, from http://cybermetrics.cindoc.csic.es/pruebas/v5i1p2.htm

Leydesdorff, L. (2003). The mutual information of university-industry-government relations: An indicator of the Triple Helix dynamics. *Scientometrics, 58*(2), 445-467.

Leydesdorff, L., & Amsterdamska, O. (1990). Dimensions of citation analysis. *Science Technology & Human Values, 15*(3), 305-335.

Leydesdorff, L., & Curran, M. (2000). Mapping University-Industry-Government relations on the Internet: The construction of Indicators for a Knowledge-Based Economy. *Cybermetrics, 4*(1). Retrieved January 7, 2005, from http://cybermetrics.cindoc.csic.es/cybermetrics/articles/v4i1p2.html

Leydesdorff, L., & Etzkowitz, H. (2003). Can "the public" be considered as a fourth helix of university-industry-government relations? *Science and Public Policy, 30*(1), 55-61.

Leydesdorff, L., & Meyer, M. (2003). The Triple Helix of university-industry-government relations: Introduction to the topical issue. *Scientometrics, 58*(2), 191-203.

Leydesdorff, L., & Vaughan, L. (2006). Co-occurrence matrices and their applications in information science: Extending ACA to the web environment. *Journal of the American Society for Information Science and Technology, 57*(12), 1616-1628.

Li, X., Thelwall, M., Musgrove, P., & Wilkinson, D. (2003). The relationship between the WIFs or inlinks of computer science departments in the UK and their RAE ratings or research productivities in 2001. *Scientometrics, 57*(2), 239-255.

Li, X. (2005). *National and International University Departmental Web Site Interlinking: A Webometric Analysis*. Ph.D Thesis, University of Wolverhampton.

Lin, D., & Loui, M. C. (1998). Taking the byte out of cookies: Privacy, consent and the web. *ACM SIGCAS Computers and Society, 28*(2), 39-51.

Live Search. (2007). *We are flattered but…….* Retrieved May 12, 2007, from http://blogs.msdn.com/livesearch/archive/2007/03/28/we-are-flattered-but.aspx

Live Search. (2007). *Advanced search keywords*. Retrieved May 2, 2007, from http://help.live.com/Help.aspx?market=en-US&project=WL_Searchv1&querytype=topic&query=WL_SEARCH_REF_AdvancedSearch.htm

Lundvall, B. A. (Ed.). (1992). *National Systems of Innovation: Towards a theory of innovation and interactive learning*. London: Pinter.

MacRoberts, M. H., & MacRoberts, B. R. (1984). The negational reference: or the art of dissembling. *Social Studies of Science, 14*(1), 91-94.

MacRoberts, M. H., & MacRoberts, B. R. (1988). Author Motivation for Not Citing Influences: A Methodological Note. *Journal of the American Society for Information Science, 39*(6), 432-433.

MacRoberts, M. H., & MacRoberts, B. R. (1989). Problems of Citation Analysis: A Critical Review. *Journal of the American Society for Information Science, 40*(5), 342-349.

Maddox, L. M. (1999). The use of pharmaceutical web sites for prescription drug information and product requests. *Journal of Product & Brand Management, 8*(6), 488-496.

Mählck, P., & Persson, O. (2000). Socio-bibliometric mapping of intradepartmental networks. *Scientometrics, 49*(1), 81-91.

Manion, M., & Goodrum, A. (2000). Terrorism or civil disobedience: Toward a hacktivist ethic. *Computers and Society, 30*(2), 14-19.

Martín-Sempere, M. J., Rey-Rocha, J., & Garzón-García, B. (2002). The effect of team consolidation on research collaboration and performance of scientists: Case study of Spanish university researchers in Geology. *Scientometrics, 55*(3), 377-394.

McMillan, G. S. (2000). Using bibliometrics to measure firm knowledge: An analysis of the US pharmaceutical industry. *Technology Analysis & Strategic Management, 12*(4), 465-475.

McMillan, S. J. (2000). The Microscope and the Moving Target: The challenge of applying content analysis to the world wide web. *Journalism & Mass Communication Quarterly, 77*(1), 80-98.

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

Mettrop, W., & Nieuwenhuysen, P. (2001). Internet search engines: Fluctuations in document accessibility. *Journal of Documentation, 57*(5), 623-651.

Meyer, M. (2000a). Does science push technology? Patents citing scientific literature. *Research Policy, 29*(3), 409-434.

Meyer, M. (2000b). What is special about patent citations? Differences between scientific and patent citations. *Scientometrics, 49*(1), 93-123.

Meyer, M., & Bhattacharya, S. (2004). Commonalities and differences between scholarly and technical collaboration: An exploration of co-invention and co-authorship analyses. *Scientometrics, 61*(3), 443-450.

Middleton, I., McConnell, M., & Davidson, G. (1999). Presenting a model for the structure and content of a university World Wide Web site. *Journal of Information Science, 25*(3), 219-227.

Mikheev, A. (2003). Text segmentation. In: R. Mitkov (ed.) *The Oxford Handbook of Computational Linguistics.* (pp.201-218). Oxford: Oxford University Press.

Moed, H. F. (2005). *Citation analysis in research evaluation*. Springer: Dordrecht.

Moravcsik, M. J., & Murugesan, P. (1975). Some results on the function and quality of citations. *Social Studies of Science, 5*(1), 86-92.

MSDN. (2006). *Search Macros: LinkfromDomain*. Retrieved November 20, 2006, from http://blogs.msdn.com/livesearch/archive/2006/10/16/search-macros-linkfromdomain.aspx

Musgrave, S. (2004). The community portal challenge – is there a technology barrier for local authorities? *Telematics and Informatics, 21*(3), 261-272.

Musgrove, P., Binns, R., Page-Kennedy, T., & Thelwall, M. (2003). A method for identifying clusters in sets of interlinking Web spaces. *Scientometrics, 58*(3), 657-672.

Narin, F. (1994). Patent bibliometrics. *Scientometrics, 30*(1), 147-155.

Narin, F., Noma, E., & Perry, R. (1987). Patents as indicators of corporate technological strength. *Research Policy, 16*(2-4), 143-155.

National Statistics. (2006). *Business internet use*. Retrieved May 12, 2007, from http://www.statistics.gov.uk/downloads/theme_economy/ecommerce_report_2005.pdf

Negroponte, N. (1995). *Being digital*. London: Coronet.

Neuendorf, K. A. (2002) *The Content Analysis Guidebook*. London: Sage Publications.

Nicholas, D., Huntington, P., & Williams, P. (2002). Evaluating metrics for comparing the use of web sites: a case study of two consumer health web sites. *Journal of Information Science, 28*(1), 63-75.

Nominet. (2006). *Second level domain names*. Retrieved March 1, 2006, from http://www.nominet.org.uk/registrants/sld/

Norris, M., & Oppenheim, C. (2003). Citation counts and the Research Assessment Exercise V: Archaeology and the 2001 RAE. *Journal of Documentation, 59*(6), 709-730.

Norris, M., & Oppenheim, C. (2007). Comparing alternatives to the Web of Science for coverage of the social sciences' literature. *Journal of Informetrics, 1*(2), 161-169.

OECD & Eurostat. (2005). *Oslo Manual: Guidelines for collecting and interpreting innovation data* (3rd ed.). Retrieved May 12, 2007, from http://epp.eurostat.ec.europa.eu/cache/ITY_PUBLIC/OSLO/EN/OSLO-EN.PDF

Olsen, S. (2003). *Google cache raises copyright concerns*. Retrieved February 23, 2005, from http://news.com.com/2100-1038_3-1024234.html

O'Malley, M. A., McOuat, G. R., & Doolittle, W. F. (2002, November). The Triple Helix account of scientific innovation: A critical appraisal. Paper presented at *4th Triple Helix*

Conference, November, Copenhagen, Denmark. Retrieved June 6, 2007, from http://www.leydesdorff.net/th4cd/th4cd.zip

Oppenheim, C. (1997). The correlation between citation counts and the 1992 research assessment exercise ratings for British research in genetics, anatomy and archaeology. *Journal of Documentation, 53*(5), 477-487.

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). New Jersey: Information Today.

Oppenheim, C., Morris, A., McKnight, C., & Lowley, S. (2000). The evaluation of WWW search engines. *Journal of Documentation, 56*(2), 190-211.

Ortega, J.L., & Aguillo, I.F. (1997). Interdisciplinary relationships in the Spanish academic web space: A webometric study through networks visualization. *Cybermetrics, 11*(1). Retrieved January, 4, 2007, from http://www.cindoc.csic.es/cybermetrics/articles/v11i1p4.htmlhtm

Otte, E. & Rousseau, R. (2002). Social network analysis: A powerful strategy, also for the information science. *Journal of Information Science, 28*(6), 441-453.

Park, H W., Barnett, G. A., & Nam, I. Y. (2002). Hyperlink-affiliation network structure of top web sites: Examining affiliates with hyperlink in Korea. *Journal of the American Society for Information Science and Technology, 53*(7), 602-611.

Park, H.W., & Thelwall, M. (2003). Hyperlink analyses of the world wide web: A review. *Journal of Computer Mediated Communication, 8*(4). Retrieved June 20, 2004, from http?//www.ascusc.org/jcmc/vol8/issue4/park.html

Patel, N. (1973). Collaboration in the professional growth of American sociology. *Social Science Information, 12*(6), 77-92.

Penslar, R. L. (Ed.). (1995). *Research ethics: Cases and materials*. Bloomington, IN: Indiana University Press.

Petricek, V., Escher, T., Cox, I. J., & Margetts, H. (2006, May). The Web Structure of E-Government – Developing a Methodology for Quantitative Evaluation. *The 15th International World Wide Web Conference, 2006, May 22-26, Edinburgh, Scotland*. Retrieved May 13, 2007, from http://www.governmentontheweb.org/downloads/papers/WWW2006-Web_Structure_of_E_Government.pdf

Potratz, W., & Widmaier, B. (1996). Industrial transformation in central and eastern Europe: Is innovation a way to integration? *MOCT-MOST, 6*(4), 55-70.

Priego, J. L. O. (2003). A Vector Space Model as a methodological approach to the Triple Helix dimensionality: A comparative study of biology and biomedicine centres of two

European national research councils from a webometric view. *Scientometrics, 58*(2), 429-443.

Prime, C., Bassecoulard, E., & Zitt, M. (2002). Co-citations and co-sitations: A cautionary view on an analogy. *Scientometrics, 54*(2), 291-308.

Pritchard, A. (1969). Statistical bibliography or bibliometrics? *Journal of Documentation, 25*(4), 348-349.

Qiu, J., Chen, J., & Zhi, W. (2003) An analysis of backlink counts and Web impact factors for Chinese university Websites. In G. Jiang, R. Rousseau, & Y. Wu (Eds.), *Proceedings of the 9th International Conference of the International Society for Scientometrics and Informetrics*, 221-229. Dalian, China: Dalian University of Technology Press.

Van Raan, A. F. J. (1998). In matters of quantitative studies of science the fault of the theorists is offering too little and asking too much. *Scientometrics, 43*(1), 129-139.

Van Raan, A. F. J. (2001). Bibliometrics and the internet: Some observations and expectations. *Scientometrics, 50*(1), 59-63.

Van Raan, A. F. J. (2004). Measuring science. In H. F. Moed, W. Glänzel, & U. Schmoch (Eds.), *Handbook of Quantitative Science and Technology Research* (pp.19-50). Dordrecht: Kluwer Academic Publishers.

RAE. (2005). *RAE 2008*. Retrieved January 7, 2005, from http://www.rae.ac.uk/default.htm

Rall, D. N. (2004a). Exploring the breadth of disciplinary backgrounds in internet scholars participating in AoIR meetings, 2000-2003. *Proceedings of AoIR 5.0, 2004, September 19-22, University of Sussex, Brighton, UK*. Retrieved February 23, 2005, from http://gsb.haifa.ac.il/~sheizaf/AOIR5/399.html

Rall, D. N. (2004b). Locating Internet research methods within five qualitative research traditions. *Proceedings of the AoIR-ASIST 2004 Workshop on Web Science Research Methods, 2004, September 19, University of Sussex, Brighton, UK*. Retrieved February 23, 2005, from http://cybermetrics.wlv.ac.uk/AoIRASIST/

Reiman, J. H. (1995). Driving to the Panopticon: A philosophical exploration of the risks to privacy posed by the highway technology of the future. *Santa Clara Computer and High Technology Law Journal, 11*(1), 27-44.

Research Councils UK. (2004). *Material world: Knowledge economy*. Retrieved May 12, 2007, from http://www.rcuk.ac.uk/cmsweb/downloads/rcuk/publications/knowledge_economy.pdf

Research Councils UK. (2006). *Research Councils UK*. Retrieved May 12, 2007, from http://www.rcuk.ac.uk

Rooksby, E. (2002). *E-mail and ethics*. London: Routledge.

Rosenberg, N. (1994). *Exploring the Black Box: Technology, Economics, and History*. New York: Cambridge University Press.

Roth, D.L. (2005). The emergence of competitors to the Science Citation Index and the Web of Science. *Current Science, 89*(9), 1531-1536.

Rousseau, R. (1997). Sitations: an exploratory study. *Cybermetrics, 1*(1). Retrieved May 4, 2006, from http://cybermetrics.cindoc.csic.es/pruebas/v1i1p1.htm

Rousseau, R. (1999). Daily time series of common single word searches in AltaVista and NorthernLight. *Cybermetrics, 2/3*(1). Retrieved June 12, 2005, from http://cybermetrics.cindoc.csic.es/pruebas/v2i1p2.htm

Schäfer, W. (1983). *Finalization in science: The social orientation of scientific progress*. Lancaster: D. Reidal.

Schneier, B. (2004). *Secrets and lies: Digital security in a networked world*. New York: Hungry Minds Inc.

Schwartz, C. (1998). Web search engines. *Journal of the American Society for Information Scien*ce, *49*(11), 973-982.

Search Engine Optimization Ethics. (2002). *SEO code of ethics*. Retrieved March 1, 2005, from http://www.searchengineethics.com/seoethics.htm

SearchEngineWatch. (2006). *Windows Live Search Adds linkfromdomain Command*. Retrieved November 15, 2006, from http://blog.searchenginewatch.com/blog/061017-094100

Shaw, D. (2001). Playing the Links: Interactivity and Stickiness in .Com and "Not.Com" web Sites. *First Monday, 6*(3). Retrieved June 12, 2005, from http://firstmonday.org/issues/issue6_3/shaw/index.html

Shaw, W. H. (1999). *Contemporary ethics: Taking account of utilitarianism*. Oxford: Blackwell.

Smith, A. (1999a). *ANZAC webometrics: Exploring Australasian Web structures*. Retrieved May 12, 2007, from http://www.csu.edu.au/special/online99/proceedings99/203b.htm

Smith, A. G. (1999b). A tale of two web spaces. *Journal of Documentation, 55*(5), 577-592.

Smith, A. G. (1999c). *The impact of web sites: a comparison between Australasia and Latin America*. Retrieved May 12, 2007, from http://www.vuw.ac.nz/staff/alastair_smith/publns/austlat/

Smith, A. (2003). Classifying links for substantive Web Impact Factors. In G. Jiang, R. Rousseau, & Y. Wu (Eds.), *Proceedings of the 9th International Conference of the International Society for Scientometrics and Informetrics* (pp.305-311). Dalian, China: Dalian University of Technology Press.

Smith, A. G. (2004). Web links as analogues of citations. *Information Research, 9*(4). Retrieved May 12, 2007, from http://informationr.net/ir/9-4/paper188.html

Smith, A.G. (2005). Citations and links as a measure of the effectiveness of online LIS journals. *IFLA Journal, 31*(1), 76-84.

Smith, A., & Thelwall, M. (2002). Web Impact Factor for Australasian universities. *Scientometrics, 54*(3), 363-380.

Smith, K. (2005). Measuring Innovation. In: J. Fagerberg, D. Mowery, R.R. Nelson (eds.) *The Oxford Handbook of Innovation.* (pp.148-177). Oxford: Oxford University Press.

Snyder, H., & Rosenbaum, H. (1999). Can search engines be used as tools for web-link analysis? A critical view. *Journal of Documentation, 55*(4), 375-384.

Statistical Cybermetrics Research Group. (2007). *The academic web link database project: Making available databases of academic web links to the world research community*. Retrieved May 12, 2007, from http://cybermetrics.wlv.ac.uk/database/index.html

Stitt, R. (2004). Curbing the Spam problem. *IEEE Computer, 37*(12), 8.

Stokes, D. E. (1997). *Pasteur's Quadrant: Basic science and technological innovation*. Washington D.C.: Brookings Institution Press.

Stuart, D., & Thelwall, M. (2005). What can university-to-government web links reveal about university-government collaborations? In P. Ingwersen, & B. Larsen (Eds.), *Proceedings of the 10th International Conference of the International Society for Scientometrics and Informetrics*: *Vol*. *1*. (pp.188-192). Stockholm: Karolinska University Press.

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.

Stuart, D., & Thelwall, M. (2007). University-industry-government relationships manifested through MSN reciprocal links. In D. Torres-Salinas, & H. F. Moed (Eds.), *Proceedings of the 11th International Conference of the International Society for Scientometrics and Informetrics*: *Vol*. *2*. (pp.731-735). Madrid: CINDOC.

Stuart, D., Thelwall, M., & Harries, G. (2007). UK academic web links and collaboration – an exploratory study. *Journal of Information Science, 33*(2), 231-246.

Subotzky, G. (1999). Alternatives to the entrepreneurial university: New modes of knowledge production in community service programs. *Higher Education, 38*(4), 401-440.

Sullivan, D. (2004). *Search engines and legal issues.* SearchEngineWatch. Retrieved February 23, 2005, from http://searchenginewatch.com/resources/article.php/2156541

Tague-Sutcliffe, J. (1992). An introduction to informetrics. *Information Processing & Management, 38*(1), 1-3.

Tang, R., & Thelwall, M. (2003a). Patterns of international and national web inlinks to US university departments: A webometric analysis of disciplinary specificity. In G. Jiang, R. Rousseau, & Y. Wu (Eds.), *Proceedings of the 9th International Conference of the International Society for Scientometrics and Informetrics* (pp.312-320). Dalian, China: Dalian University of Technology Press.

Tang, R., & Thelwall, M. (2003b). U.S. academic departmental web-site interlinking in the United States: Disciplinary differences. *Library & Information Science Research*, *25*(4), 437-458.

Tang, R., & Thelwall, M. (forthcoming). A hyperlink analysis of US public and academic libraries' Web sites, *Library Quarterly*.

Terveen, L., & Hill, W. (1998) Evaluating emergent collaboration on the web. In S. Poltrock, & J. Grudin (Eds.), *Proceedings of the 1998 ACM conference on computer supported cooperative work* (pp.355-362). ACM Press.

Thelwall, M. (2000). Web Impact Factors and Search Engine Coverage, *Journal of Documentation*, *56*(2), 185-189.

Thelwall, M. (2001a). A Web Crawler Design for Data Mining. *Journal of Information Science*, *27*(5), 319-325.

Thelwall, M. (2001b). Commercial web site links. *Internet Research*, *11*(2), 114-124.

Thelwall, M. (2001c). Exploring the link structure of the Web with network diagrams. *Journal of Information Science*, *27*(6), 393-401.

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

Thelwall, M. (2001e). Results from a Web Impact Factor crawler. *Journal of Documentation*, *57*(2), 177-191.

Thelwall, M. (2001f). The responsiveness of search engine indexes. *Cybermetrics*, *5*(1). Retrieved June 12, 2005, from http://cybermetrics.cindoc.csic.es/pruebas/v5i1p1.htm

Thelwall, M. (2002a). A comparison of sources of links for academic Web impact factor calculations. *Journal of Documentation*, *58*(1), 66-78.

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

Thelwall, M. (2002c). A research and institutional size based model for national university web site interlinking. *Journal of Documentation*, *58*(6), 683-694.

Thelwall, M. (2002d). Conceptualizing the documentations on the web: An evaluation of different heuristic-based models for counting links between university web sites. *Journal of the American Society for Information Science and Technology*, *53*(12), 995-1005.

Thelwall, M. (2002e). Evidence for the existence of geographic trends in university Web site interlinking. *Journal of Documentation, 58*(5), 563-574.

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

Thelwall, M. (2002g). Research dissemination and invocation on the web. *Online Information Review,* 26(6), 413-420.

Thelwall, M. (2002h). 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. (2003a). A free database of University Web Links: Data collection issues. *Cybermetrics*, *6/7*(1). Retrieved January 7, 2005, from http://cybermetrics.cindoc.csic.es/cybermetrics/articles/v6i1p2.html.

Thelwall, M. (2003b). A layered approach for investigating the topological structure of communities in the web. *Journal of Documentation*, *59*(4), 410-429.

Thelwall, M. (2003c). Web use and peer interconnectivity metrics for academic web sites. *Journal of Information Science, 29*(1), 1-10.

Thelwall, M. (2003d). What is this link doing here? Beginning a fine-grained process of identifying reasons for academic hyperlink creation. *Information Research*, *8*(3). Retrieved January 7, 2005, from http://informationr.net/ir/8-3/paper151.html.

Thelwall, M. (2004a). *Link Analysis: An Information Science Approach*. San Diego, CA: Academic Press.

Thelwall, M. (2004b). Methods for reporting on the targets of links from national systems of university Web sites. *Information Processing & Management, 40*(1), 125-144.

Thelwall, M. (2004c). Weak benchmarking indicators for formative and semi-evaluative assessment of research. *Research Evaluation*, *13*(1), 63-68.

Thelwall, M. (2005). Text characteristics of English language university web sites. *Journal of the American Society for Information Science and Technology, 56*(6), 609-619.

Thelwall, M. (forthcoming). Extracting Accurate and Complete Results from Search Engines: Case Study Windows Live.

Thelwall, M., Binns, R., Harries, G., Page-Kennedy, T., Price E., & Wilkinson, D. (2002). European Union Associated University Websites. *Scientometrics, 53*(1), 95-111.

Thelwall, M., & Harries, G. (2003). The Connection between the Research of a University and Counts of Links to its Web Pages: An Investigation Based Upon a Classification of the Relationships of Pages to the Research of the Host University. *Journal of the American Society for Information Science and Technology, 54*(7), 594-602.

Thelwall, M., & Harries, G. (2004a). Can personal web pages that link to universities yield information about the wider dissemination of research? *Journal of Information Science*, *30*(3), 243-256.

Thelwall, M., & Harries, G. (2004b). Do better scholars have significantly higher online impact? *Journal of the American Society for Information Science and Technology*, *55*(2), 149-159.

Thelwall, M., Harries, G., & Wilkinson, D. (2003). Why do web sites from different academic subjects interlink? *Journal of Information Science, 29*(6), 453-471.

Thelwall, M., & Price, L. (2003). Disciplinary difference in academic web presence: A statistical study of the UK. *Libri, 53(*4), 242-253.

Thelwall, M., & Price, L. (2006). Language evolution and the spread of ideas on the Web: A procedure for identifying emergent hybrid word family members. *Journal of the American Society for Information Science and Technology*, *57*(10), 1326-1337.

Thelwall, M., & Price, L. (Forthcoming). A generic lexical framework for link list and URL list analysis. *Journal of the American Society for Information Science and Technology.*

Thelwall, M., & Smith, A. (2002). Interlinking between Asia-Pacific university web sites. *Scientometrics, 55*(3), 363-376.

Thelwall, M., & Stuart, D. (2006). Web crawling ethics revisited: Cost, privacy, and denial of service. *Journal of the American Society for Information Science and Technolohgy*, *57*(13), 1771-1779.

Thelwall, M., & Tang, R. (2003a). Disciplinary and linguistic considerations for academic Web linking: An exploratory hyperlink mediated study with Mainland China and Taiwan. *Scientometrics, 58*(1), 153-179.

Thelwall, M., Tang, R., & Price, L. (2003). Linguistic patterns of academic web use in Western Europe. *Scientometrics, 56*(3), 417-432.

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

Thelwall, M., & Wilkinson, D. (2003b). Three target document range metrics for university Web sites. *Journal of the American Society for Information Science and Technology*, *54*(6), 489-496.

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

Thomas, O., & Willett, P. (2000). Webometric analysis of departments of librarianship and information science. *Journal of Information Science, 26*(6), 421-428.

Underwood, J. (2001). *Competitive intelligence*. New York: Capstone Express Exec, Wiley.

University of Wolverhampton. (2005). *UK University Names and Domain Names.* Retrieved June 12, 2005, from http://cybermetrics.wlv.ac.uk/database/university_lists.htm

USPTO (2007). Assignee Name (AN). http://www.uspto.gov/patft/help/helpflds.htm#Assignee_Name

Van Couvering, E. (2004). New media? The political economy of Internet search engines. *Paper presented at the Annual Conference of the International Association of Media & Communications Researchers, 2004, July 24, Porto Alegre, Brazil.*

Van Looy, B., Callaert, J., & Debackere, K. (2006). Publication and patent behaviour of academic researchers: Conflicting, reinforcing or merely co-existing? *Research Policy*, *35*(4), 596-608.

Van Looy, B., Ranga, M., Callaert, J., Debackere, K., & Zimmermann, E. (2004). Combining entrepreneurial and scientific performance in academia: towards a compounded and reciprocal Matthew-effect? *Research Policy*, *33*(3), 425-441.

Vardy, P., & Grosch, P. (1999). *The puzzle of ethics*. London: Fount.

Vasileiadou, E., & van den Besselaar, P. (2006). Linking shallow, linking deep. How scientific intermediaries use the Web for their network of collaborators. *Cybermetrics, 10*(1). Retrieved September 3, 2006, from www.cindoc.csic.es/cybermetrics/articles/v10i1p5.html

Vaughan, L. (2004a). Exploring website features for business information. *Scientometrics*, *61*(3), 467-477.

Vaughan, L. (2004b). New measurements for search engine evaluation proposed and tested. *Information Processing and Management*, *40*(4), 677-691.

Vaughan, L., & Shaw, D. (2003). Bibliographic and web citations: What is the difference? *Journal of the American Society for Information Science and Technology, 54*(14), 1313-1322.

Vaughan, L. & Shaw, D. (2005). Web citation data for impact assessment: A comparison of four science disciplines. *Journal of the American Society for Information Science and Technology, 56*(10), 1075-1087.

Vaughan, L., & Thelwall, M. (2003a). Scholarly use of the web: What are the key inducers of links to journal web sites? *Journal of the American Society for Information Science and Technology, 54*(1), 29-38.

Vaughan, L., & Thelwall, M. (2003b). Web link counts correlate with ISI impact facors: Evidence from two disciplines. *Journal of the American Society for Information Science and Technology, 54*(1), 29-38.

Vaughan, L., & Thelwall, M. (2004). Search engine coverage bias: evidence and possible causes. *Information Processing & Management, 40*(4), 693-707.

Vaughan, L., & Wu. G. (2004). Links to commercial websites as a source of business information. *Scientometrics, 60*(3), 487-496.

Vinkler, P. (1994). The origin and features of information referenced in pharmaceutical patents. *Scientometrics, 30*(1), 283-302.

Wasserman, S., & Faust, K. (1994). *Social Network Analysis: Methods and applications.* CUP: Cambridge.

Webster, A., & Packer, K. (2001). When worlds collide: Patents in public-sector research. In H. Etzkowitz, & L. Leydesdorff (Eds.), *Universities and the Global Knowledge Economy: A Triple Helix of University-Industry-Government Relations* (pp.47-59). London: Continuum.

Weingart, P. (1997). From "Finalization" to "Mode 2": old wine in new bottles? *Social Science Information, 36*(4), 591-613.

West Midlands Higher Education Association. (2005). *The West Midlands Higher Education Association.* Retrieved June 12, 2005, from http://www.wmhea.ac.uk

White, H. (2001). Authors as citers over time. *Journal of the American Society for Information Science and Technology, 52*(2), 87-108.

White, H. D., Wellman, B., & Nazer, N. (2004). Does citation reflect social structure? Longitudinal evidence from the "Globenet" interdisciplinary research group. *Journal of the American Society for Information Science and Technology, 55*(2), 111-126.

Whitley, R. (2000). *The Intellectual and Social Organization of the Sciences* (2nd ed.). Oxford: Oxford University Press.

Wilcox, R. R. (2003). *Applying Contemporary Statistical Techniques*. London: Elsevier Science.

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.

Wilkinson, D., Thelwall, M., & Li, X. (2003). Exploiting hyperlinks to study academic Web use. *Social Science Computer Review, 21*(3), 340-351.

Williamson, E., & Smyth M. (Eds.). (2004). *Researchers and their subjects, ethics, power, knowledge and consent*. Bristol: Policy Press.

Wikia Search. (2007). *Welcome to Wikia Search*: *A project by Wikia to create the search engine that changes everything*. Retrieved May 10, 2007, from http://search.wikia.com/wiki/Search_Wikia

Wormell, I. (2001). Informetrics and Webometrics for measuring impact, visibility, and connectivity in science, politics and business. *Competitive Intelligence Review*, *12*(1), 12-23.

Wronkiewicz, K. (1997). Spam like fax. *Internet World, 8*(2), 10.

Yahoo! (2005). *How can I reduce the number of requests you make on my web site?* Yahoo inc. Retrieved February 23, 2005, from http://help.yahoo.com/help/us/ysearch/slurp/slurp-03.html

Ziman, J. (1994). *Prometheus Bound: Science in a dynamic steady state.* Cambridge: Cambridge University Press.

Appendix 1 - Classification protocol for determining the reason for link placement

(1) Reflecting a collaborative relationship
(a) Research relationship/collaboration
The link must be embedded either on a page, or an area of a page that is referring to a piece of research, or a research group.
And
Uses terminology such as: partners; run by; in collaboration with; or near synonyms with reference to the organisation/group/person represented by the link.

(b) Business client
(i) It is stated that work has been carried out previously for the organisation/group/person represented by the link.
(ii) It is stated that techniques/technologies developed by the source page owner are or have been used by the organisation/group/person represented by the link.

(c) Previous employer
It is stated that at least one person involved with the source page has worked/been employed by the organisation/group/person represented by the link.

(d) Sponsorship/Funding
It is stated that money has been supplied to fund a: post, studentship, project or event.

(e) Shared Interests
The link must be embedded on a page, or an area of the page that is not referring to a piece of research.
And
Uses terminology to imply an activity where the creators of the source page work with the organisation/group/person represented by the link, but one is not deemed to be the client of the other.

(f) Reflecting membership of an organisation
It is stated that someone represented by the source page is a member of another organisation/group represented by the link.

(g) Informal relationship
The text around the link indicates that the organisation/group/person represented by the link has participated with the source page, without working directly with the source page owner, e.g., attended a conference, given a lecture etc.
(h) Contractor
Target page owner has provided work for source page owner.

(i) Undisclosed relationship
Either

(i) A relationship is inferred by the placement of an isolated link (nothing to indicate it's purpose) away from the main body of the page. There may be a number of links together (either as text or as graphics), but they must then form a banner at either the top (above or below the title) or the bottom (above or below any credits) of the page.
A relationship cannot be inferred from an isolated link if the link is to a recognizable product to aid internet use, e.g., search engines or media players.
Or
(ii) Use of vague terminology with reference to the link (such as 'supported by') which whilst implying a relationship does not allow the link to be classified under one of the categories (a)-(h).

(2) Non-collaborative relationship
The page is available to be viewed, the link is still on the page, but the link does not fulfil one of the criteria necessary to be classified as reflecting a relationship.

(3) Unable to determine the reason
Either the link is no longer on the page, or the page is no longer available.