The World Wide Web of Science: Reconfiguring Access to Information

Ralph Schroeder, Alexandre Caldas, Gustavo Mesch, William Dutton

Oxford Internet Institute, Oxford University, UK

Email address of corresponding author: Ralph Schroeder@oii.ox.ac.uk

Abstract. This paper presents preliminary results from a new study of how the Internet and the Web might reconfigure access to scientific information. The study combines qualitative and quantitative methods – in-depth interviews and webmetric analysis – to explore how the Internet and Web are reinforcing the role of existing sources of information, or tending to either 'democratize' or centralize patterns of access conforming to the expectations of a 'winner-take-all' process of selection. This paper reports the early findings of two case studies focused on the global issues of (1) climate change and (2) the Internet and society. The preliminary analyses provide some support for all three patterns – reinforcing, democratizing, and 'winner-take-all' - but also point to the need for indicators over longer periods of time and the triangulation of methods from webmetric analysis with expert groups and in-depth case studies of issue areas.

Background and Research Questions

The rapid diffusion of the Internet and World Wide Web, accompanied by the growth of online resources and information exchange, prompts the question: "To what extent is the Internet reshaping access to knowledge and science resources worldwide?"(Dutton et al., 2003). This is of particular importance in the case of science, where the Internet has become a major medium for access to information and collaboration among scholars. There are at least three competing views on how the Internet and Web might reconfigure access.

First, it is often hypothesized that the Internet will widen access to a global body of knowledge and expertise since it enables researchers to scan the world for information without increased costs. For example, Drori, Meyer, Ramirez and Schofer (2003) have identified the globalization of scientific and research institutions as a major social trend over the course of the 20th century. They argue that major scientific institutions – professional and standards bodies, education and government research agencies, and state-organized education systems – have increasingly become similarly organized world-wide, thus also undermining local and parochial patterns of communication and reinforcing a more globalized scientific research community. The Internet and Web could in this way be used so as to further the process described by Drori, Meyer, Ramirez and Schofer and others.

Alternatively, electronic resources might be used in ways that reinforce the prominence of the strongest centres of information and exisiting networks of communication, such as the major research centres and the scholars connected with them. If the Internet and Web for scientific research become a fragmented and highly stratified resource, with only a few sites used by a limited group of researchers, then this would cast doubt on the globalization perspective and point instead towards a hierarchical online order of knowledge that reinforces historically and geographically grounded networks of communication and information exchange.

A third possible outcome is captured by the concept of a 'winner-takes-all' effect (Frank and Cook, 1995) in scientific knowledge. This idea builds on a pattern identified earlier in the sociology of science as the "Matthew effect", which describes a process whereby leaders gain a cumulative advantage over time (see Merton 1988; Caldas, 2004: 107-114). According to this view, some researchers with an initial advantage obtain ever greater advantages in the reputation of their research over time, such as when their students cite them, and follow their research traditions. One question is whether the Internet and Web will facilitate and reinforce a 'winner-take-all' centralizing tendency, reinforce existing structures of influence, or democratize – further decentralize – access to scientific sources of information?

Approach

To investigate this topic, we chose a number of global research topics. The selection of global issues was partly based on a list recently put together by a panel of experts as part of a UN assessment of issues which represent top priorities among the challenges facing the world today, including climate change, HIV/AIDS, water and sanitation, governance, and trade reform. We also included some equally global but more topical and widely known issues such as 'terrorism' and 'Internet and society'. In this preliminary study, we report initial findings on two of these topics, 'climate change' and 'Internet and society'.

Together, the topics selected represent urgent issues that are arguably equally relevant in any part of the world. They also represent a mixture addressed by both natural and social sciences. And finally, they represent issues for which many leading institutions have established online resources and which have become subjects for lively online discussion groups. A number of academic centres have provided online resources dedicated to these issues, such as project pages, pages with links and data, and online discussion newsgroups and other discussion for a on these topics have proliferated rapidly in recent years (Borgmann, 2000; Nentwich, 2003).

A focus on practice is also central to this study. In science and technology studies, a common criticism of ideas about the 'impact' of new technologies is that they extrapolate from the features of technologies and do not take the actual uses of technologies into account (Edgerton, 1998). Thus the Internet and Web have been heralded as bringing about an information revolution, but have these changes been realized in practice?

Finally, the study seeks to triangulate on patterns of access through the use of Webmetrics, interviews and case studies of issue areas. For *offline* resources, it is difficult to gauge the extent to which they are *used*. Even when it is possible to

establish how widely they are distributed (number of accessible journals, memberships of professional bodies, frequency and attendance at scientific conferences and workshops), it is difficult to measure *accessing* knowledge. Online resources are different: access to websites and participation in online discussion groups can be captured and analysed electronically, utilizing Webmetrics, yielding data that reveals the geographical distribution and frequency of electronic access.

Webmetric Analyses

The Web can be regarded as a socio-technical system with a particular network structure. It is a large-scale structure that can be modeled as a network with certain properties. There is an extensive literature which provides theoretical frameworks, methods and techniques for network analysis (for example, Wasserman and Faust, 1994). These include the size of the network, connectivity, density, and other properties. Webmetric analysis has focused in particular on the relationship between inbound and outbound links, the links *directed at* a certain node and the links *originating in* a certain node (a node in this case being simply the main website, such as the homepage of a research institution, to which other pages are linked).

Much discussion in webmetric analysis has been devoted to whether the Web, as a whole or in part, follows a 'power law' distribution; that is, a mathematical formula that expresses the 'winner-take-all' hypothesis or the idea that some sites are exponentially better connected within the network. Specifically in relation to the central question in our research, Pennock et al. (2002) have argued against Barabasi and Albert (1999) that 'Winners don't take all': The 'winner take all' hypothesis, they say, may apply to the Web as a whole, but it does not apply, according to their webmetric analysis of connectivity, to university homepages for example, which exhibit a more uniform (ie. not 'rich get richer') pattern of connectivity to other university homepages.

Interviews and Case Studies

Several studies have used Webmetrics in particular to analyze the links between websites of researchers and of research institutions. However, the links between the websites of researchers or research institutions may or may not give an accurate indication of how central a researcher or institution is in the field. In this research project, we have therefore combined (quantitative) webmetric analysis with (qualitative) interviews with researchers, which also provide information about how these researchers make use of online resources. In later stages of the research, we will move toward case studies of selected research topics as a further means for interpreting the patterns we uncover.

Before presenting overall patterns identified through the Webmetric analyses, it is useful to present brief 'mini-cases' of how researchers in our two areas describe their use of the Web. These mini-cases are not randomly selected or representative, but they provide a basis for introducing the more systematic data analyses of Web linkages discussed in the next section.

Climate Change and Internet and Society Researchers

An initial exploratory phase of our research has involved interviews with several researchers, in person and via email, within two of our topic areas. We used semi-structured interviewing to span a range of issues about their use of online resources, including the websites they most frequently accessed, their participation in electronic mailing lists and discussion groups, and the kinds of information they used the Web and Internet to obtain. We present a few of these interviews with a summary of their answers about how they use online resources; the most frequently used sites and types of information sought; and how this helps in their day-to-day research.

'Internet and Society' Researchers

A UK Case

CA is a young UK researcher in the field of 'Internet and society', with a particular interest in policy issues around the governance of the internet. He has been involved in this area for more than five years and is a highly sophisticated user of the internet for research. He has his own webpage (though not a blog) in which he comments on ongoing events in his area of research and he also contributes to several electronic mailing lists as well as running one of his own.

The main websites that he uses in his research are Google, the *New York Times* technology section, and a number of sites closely related to his research topic (Larry Lessig, Creative Commons, ITU policy blog). He consults these on a daily or weekly basis. The way he identifies whether a site contains high quality information is by 'context' - in other words, the reputation of the source. Most of the sites he frequents contain material by people that are personally known to him.

This researcher says that online information has become much more important to him than offline information. He tends to seek out individual scholars rather than institutions when he tries to keep up to date with novel research, and uses a few key websites (especially blogs) as well as electronic mailing lists on a daily basis.

A Swedish Case

LD is a Swedish PhD student who has been doing research in the area of 'Internet and Society' for three years, in this case with special reference to open source software and innovation. He uses the Internet and Web on a daily basis to find data and to communicate with others in his field. His daily uses also include checking three or four mailing lists closely related to his topic and visits to several blogs.

Google and Scholar Google are by far the most frequently used sites which he uses many times every day. Next are daily checks of electronic journals and data bases and the MIT open source site. He also occasionally visits the Web pages of individual scholars' who are known to him. The daily Google searches are almost exclusively done using individual's names. He rarely uses key words and he never searches for

institutions. Scholar Google is sometimes used to refine the search by following up citation lists and special topic keywords.

LD uses online resources much more intensively than offline ones. Keeping up with individual scholars, being able to download their papers from their sites and from electronic journals, as well as keeping up with technology in newspapers and blogs are all essential daily activities.

'Climate Change' Researchers

A UK Case

MP is a young climate change researcher in the UK with more than five years experience in this field. He uses the Internet and Web to find data sets (including being able ask questions about this data) about temperature patterns and related data world-wide. He also uses the Web to find publications using the Science Citation index and to publish his research results on his webpage. In addition, he is a member of a number of mailing lists and fora, though passively rather than an as active participant.

The main websites he uses are Google and Scholar Google, BBC weather, and the journals *Nature* and *Science*, which provide him with abstracts of and access to the latest papers in his field. He checks these sites at least weekly. He uses Google with key words and to take him to institutions in his field and Scholar Google to take him to persons in his field. Typical sites of institutions where he keeps up to date with research are the National Center for Atmospheric Research in the US and the British Atmospheric Data Centre.

For MP, the key role of the Web is to access sites with data. This means that he does not need to depend on individuals far and wide to get hold of data (though in some cases he needs to ask individuals for web authorization for access to data on particular sites, which is usually obtained). Further, the Web is a source for obtaining pre-publication papers as well as project descriptions – again, without needing to ask people for these. All in all, the Web has become an indispensable to MP as a way of keeping abreast with research and obtaining data.

A Swedish Case

TB is a Swedish PhD student researching climate change who has been working in the field for more than five years, particularly in relation to carbon emissions policy. Like MP, he uses data bases on the Web extensively as well as tools like the Science Citation Index and Elsevier's Science Direct. He also subscribes to mailing lists, for example the Indian Centre for Science and Environment and also electronic fora like Future International Action on Climate Change.

There are several websites that he consults at least on a weekly basis, including from the Indian Center for Science and Environment, the Centre for International Climate and Environmental Research in Oslo, the Austrian International Institute for Applied Systems Analysis and Point Carbon which forecasts carbon emissions markets. The most authoritative site for his research is that of the International Panel on Climate Change, but he needs to consult this less frequently (monthly). TP also uses Google

on a daily basis, combining the search term ‘climate change’ or similar with the name of a person or research institution (academic, governmental, or non-governmental organization) to search for news of a particular project or the reports which are the basis for scientific journal publications.

Although TP knows several of the researchers whose work contributes to the sites that he uses, it is the project descriptions, papers and data that are provided on the institutional websites that constitute the most valuable online resources for him. Online material is essential for him to keep up to date with new data and new material on policy, though older material about climate change policy is also available in print form. Timeliness is a critical factor motivating his use of the Web.

Webmetric Analysis

Does the winner-take-all hypothesis apply to patterns of access to on the ‘Internet and society’ and ‘climate change’? Our basic approach to this question has been to develop a set of keywords for each topic and crawl the Web with a combination of search engines for these keywords. In each case we used a set of keywords for Internet and society and for climate change. These were searched by means of 31 search engines (for example Google, Lycos, Yahoo).

This method was used to generate a ‘vicinity graph’ showing the nodes and links between the sites containing these keywords. The vicinity graph allows us to understand some of the network’s structural properties, such as:

a) the overall connectivity of the graph, which indicates the cohesiveness of the whole web structure;
b) the identification of nodes in the graph with different degrees of centrality and/or connectivity; and
c) the identification of nodes in the graph with different degrees of connectivity, which can be used to calculate, for instance, the *flow* between nodes which indicate the importance of each node with regard to the flow of information through the whole network.

There are limitations to this method of analysis. One is that the choice of keywords could skew results, and therefore must be validated, for example by an expert panel which might decide to include ‘global warming’ under ‘climate change’ but not ‘El Nino’. Another problem is that the sites found under ‘climate change’ might be far removed from scientific research, such as including air-conditioner firms. This was addressed in part by selecting ‘good URLs’ from within our sample based on the occurrence of our keywords in the URL designation itself, or the title, or the keyword section of the site. Still, this method cannot avoid ‘noise’ in the data. Finally, an institution’s or person’s ‘connectivity’ on the Web may or may not map well onto their status as a well-connected institution or person. This would also require more in-depth study of the area, which we have yet to conduct.

Webmetric Analysis of ‘Climate Change’

Using only four key words in the first instance - *Climate Change, Climate Changes, Ozone Depletion and Global Warming* – yielded an initial list of 3,594 web links collected from 31 search-engines. In order to make the data more manageable, a subsample of 1,156 web references was randomly drawn and from these, in turn, a reduced subsample of 150 “good” URLs was selected that was based, as mentioned

earlier, on the occurrence of the keywork in the URL itself, in the title or the key word section. A webcrawl of each of the 150 URLs collected a total of 3,489 nodes and 20,839 links which constitute the webspace of the initial 150 nodes. This method yields the following image of the network of links relating to 'climate change':

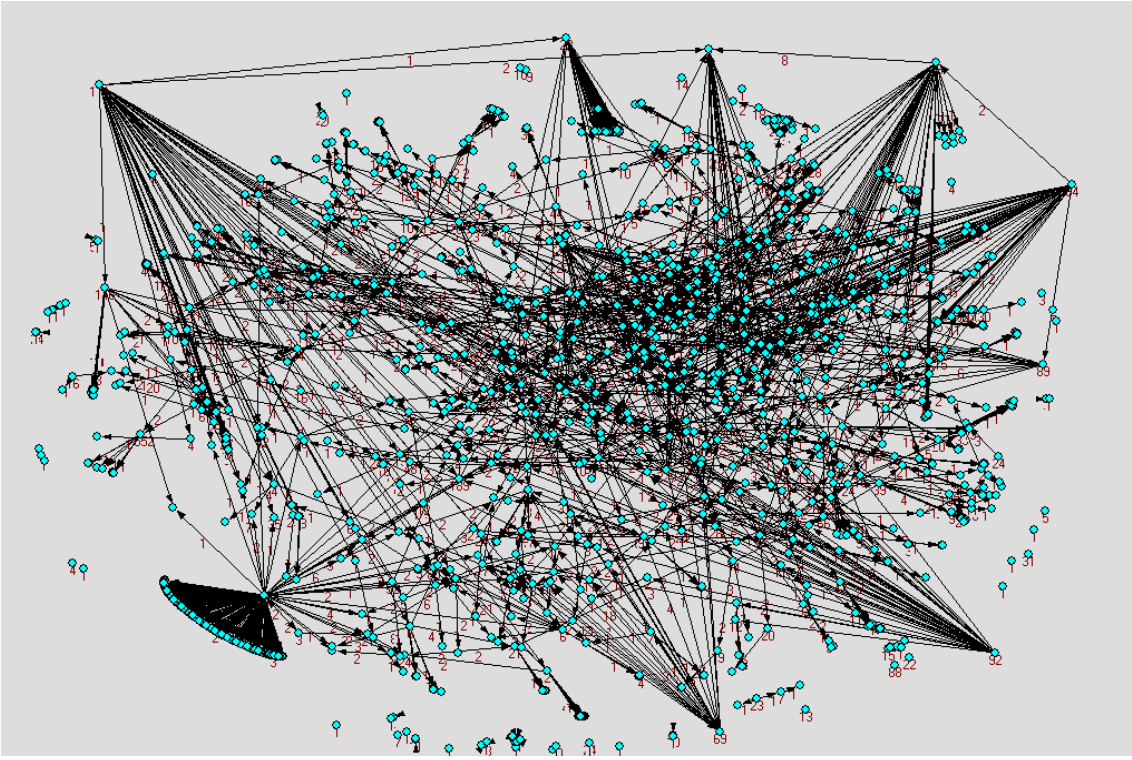

Figure 1. Web space graph of 150 URLs about 'Climate Change'

The graph shows the size of the network (nodes and links), centrality, connectivity and cohesive subgroups (or clusters). It is noteworthy that some of the sites mentioned by our interviewees, such as the International Panel on Climate change (ippc.org), are among our 150 'good' sites.

Webmetric Analysis of 'Internet and Society'

The same approach was taken in the analysis of patterns of connectivity within the area of the Internet and society. This was chosen in part because it was an area that the researchers were most familiar with, and could therefore better judge the face validity of the results. Figure 2 shows the Web space graph of 150 URLs sampled on 'Internet and Society' sites. The 150 URLs were crawled and collected a total of 3,815 nodes and 31,736 links. In a similar way to 'Climate Change' it is apparent from this representation that a "fractal-like" or highly clustered structure emerges. Again, some among these 150 sites coincided with the sites mentioned by our interviewees (Lawrence Lessig's site, the Creative Commons site).

There is a reasonable indication, corroborating previous research that a power-law distribution characterises the connectivity of these web networks. In fact, the data in figure 3 reflect a *power law* like distribution in which concerns to the *Outdegrees* for web sites on Internet and Society. This indicates a very unequal distribution of

connectivity, but does not on its own provide an indication of the "winner take all" phenomenon unless other more detailed analysis (survey and interviews and case-study research) triangulate the webmetric results. (A very similar power law distribution applies to the Climate Change network, though we do not include this figure for reasons of space.)

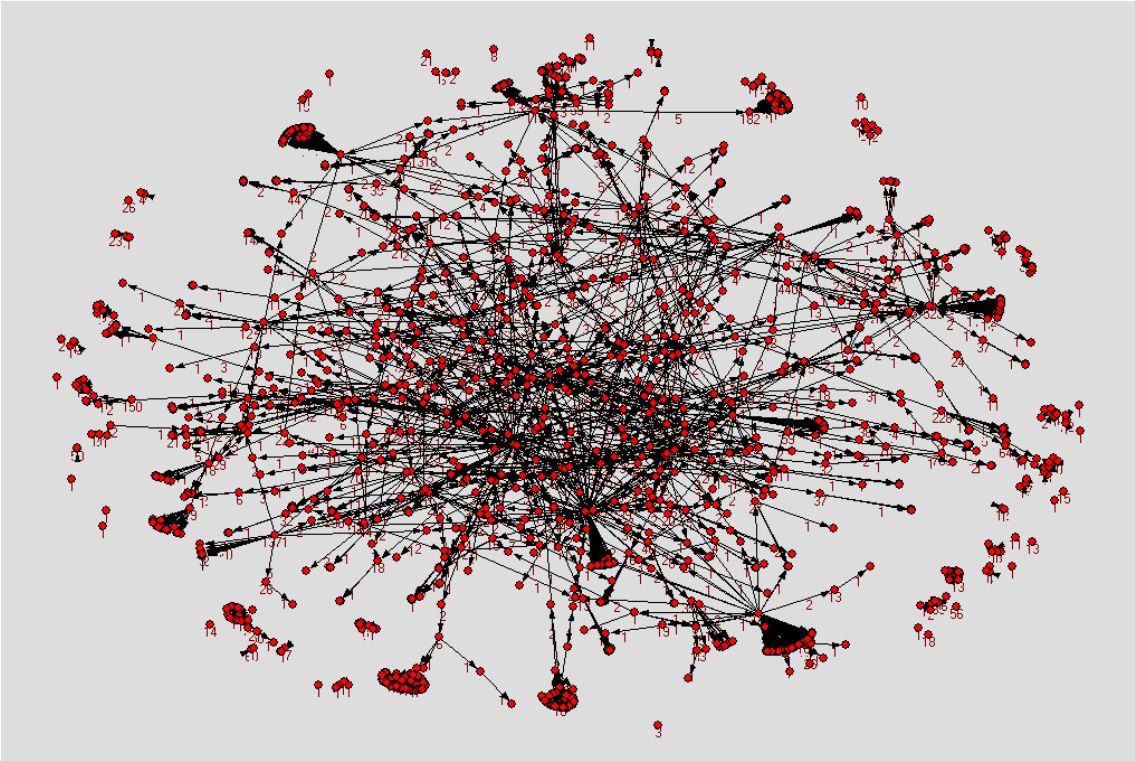

Figure 2. Webmetric Results for "Internet and Society"

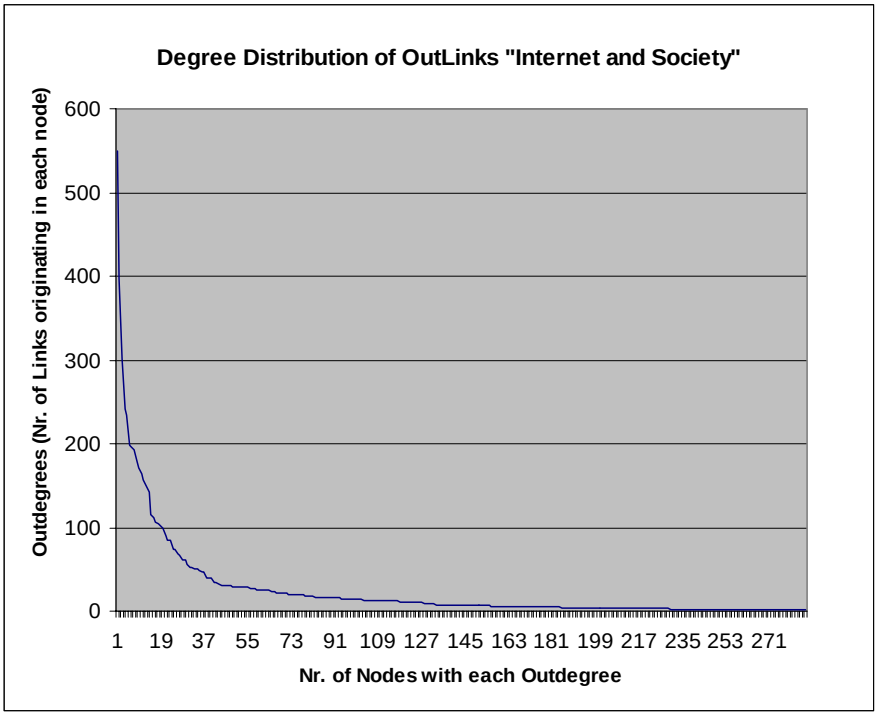

Figure 3. Outdegree Distritution for Internet and Society Network

The data in Box 1 provides additional information which allows a more complex analysis of the two networks. It turns out that the average distance between reachable pairs of nodes is quite small as compared to a random network (2.321 for climate change and 2.265 for internet and society). Both networks also show a low value for the density of the two networks (compared with the total number of potential links), but at the same time they show high levels of clusterability whereby a small number of sub-groups form 'cliques'.

	Climate Change	Internet and Society
Average distance (among reachable pairs):	2.321	2.265
Nodes in Reachable pairs:	217	295
For each pair of nodes, this indicator measures the number of edges in the shortest path between them.		
Density / average value within blocks:	0.0006	0.0005
Standard Deviations within blocks:	0.0776	0.0662
Density is a ratio of the total number of existing links as compared to the total number of potential links among nodes in the network.		
Clustering		
Number of cliques found:	190	164
A clique is defined as a subset of the network with at least 3 nodes interlinked with each other.		

Box 1. Average Distance, Connectivity and Clustering from network analysis for "Climate Change" and "Internet and Society"

In short, this additional information could point to 'democratization' (low density) and 'reinforcement' ('cliques') effects rather than a 'power law/winner-take-all' effect, but again, without additional results for other networks, longer periods of study, and triangulation of results by other methods, these are merely pointers that need to be investigated more closely.

Discussion

The Webmetric analyses proved heuristically valuable in our initial research on the World Wide Web of Science. The comparable findings across two very different areas of inquiry - Internet and Society versus Climate Change – make these results even more suggestive of an underlying process that might be common across research areas. However, these Webmetric results must be explored in other topic areas and further analysed by means of a qualitative assessment of the webmetric maps and the structural characteristics of these electronic networks. More qualitative research gained through interviews and case studies are needed to interpret the meaning and provide validation for the preliminary Webmetric results.

Ultimately, it may also be possible to make connections between the sites mentioned by our interviewees and their position in the overall network of sites analyzed through Webmetrics. The interviews and webmetric analysis that we have already conducted,

however, have been very useful in pointing to a number of problems and limitations in our research, and thus perhaps also point towards ways of overcoming them.

The early stages of a small-scale exploratory research project have a number of limitations. It is restricted to a one-point-in-time analysis of a small set of issues, for example, and the analysis is limited to English language research. However, these limitations should not prevent us from making a start on exploring the use of new information resources for research that will become increasingly vital to researchers and a public engaged in debate with researchers about global issues.

Apart from this general limitation, one problem evident from our interviews is that electronic resources are used in combination with each other, which makes it difficult to disaggregate them and specify exactly which resource is being used. An important example is the use of search engines, and in particular the widespread use of google and google.scholar. Several researchers told us that they use search engines on a daily basis and more than any other tool in searching for information. This means that from the point of view of future research, one area to investigate is how these search engines are used and what search results are obtained - in other words, how search engines act as gatekeepers to information. It is possible, for example, that any 'winner-take-all' or other effect that is uncovered by our research is in part a product of the role search engines play in the process. Research might therefore focus on the extent to which search engines are used compared with other tools.

The use of electronic resources in combination applies not only to search engines. Researchers who use electronic databases or electronic fora, for example, often use them in order to find links to other online sources. One researcher told us, for example, that instead of receiving a lot of emails from an electronic mailing list, he/she instead visited the list's website and searched the archive from this site at his/her convenience.

In order to get a sense of the most important electronic resources that scholars use, it is therefore necessary to obtain a 'holistic' or comprehensive picture of various uses of electronic resources and how they are used in combination, all within specific subject domains. Such a comprehensive picture could be obtained by means of interviewing, but another powerful technique might be to log all of a researcher's activities in order to arrive at an in-depth understanding of the combination of tools that are used.

A limitation of our webmetric analysis lies in inferring the structure of the 'whole network' based on crawling a limited number of institutional websites. There is bound to be a significant number of sites not covered by the selection process, which undermines the extent to which the sample is representative of the whole web. Nevertheless, working with a large number of sites mitigates this problem to a significant degree.

In terms of asking our interviewees directly about the 'winner-take-all' effect of electronic resources on their work, this is problematic. When we do this, our informants say that it has become easier to communicate and collaborate with far-flung colleagues as well as easier to search for and obtain information from far-flung and dispersed sources. In this sense, they explicitly support the hypothesis that online resources have enabled 'more dispersed' ways of working and sharing information.

Their answers about whether online knowledge in general is becoming more centralized or the opposite, however, are more elusive and typically of the 'it could go either way' or 'don't know/not sure' kind. This is not surprising since researchers feel able to comment on what they know (their own behaviour) but unable to comment on something that they have no direct experience of or knowledge about, viz. what other scholars may be doing, or how their domain of online expertise is structured apart from their own uses of it.

It seems from our interviews that researchers nowadays have shifted much of their activity of keeping up-to-date with research online. Several told us that they rarely use libraries or seek out offline copies of journals; this is too costly in terms of time and effort (and perhaps money for photocopying). They use electronic resources instead. However, it is difficult to determine how exclusive their use of online as opposed to offline resources is because in many cases, the same resources are available in online and offline formats. A number of other issues have been mentioned which will need to be resolved in order to get a firmer grasp of the implications of using online sources of expertise.

Conclusion

Our research has only begun to scratch the surface of evidence concerning how the Internet and Web might reconfigure access to information. Evidence of the power law operating on the Web provides some support for a 'winner-take-all' effect. However, the clustering of nodes within both topic areas is not consistent with a winner-take-all. Instead, it could suggest a reinforcement of existing networks of communication and research. Alternatively, it might represent a winner-take-all process within more specifically defined research areas. Finally, proponents of the globalizing and democratizing impact of the Internet and the Web might find evidence in the sheer size and scale – and low density - of the global networks of information exchange identified by our Webmetric analyses. Much more research remains to be done in order to weigh the relative explanatory power of these different models. But it is clear that the increasing use of online resources has important implications for gathering information. Further research will show whether and to what extent the three effects – reinforcement, democratization, or winner-take-all – apply uniformly or in different ways to scientific information.

Acknowledgments

The authors acknowledge the ESRC 'Science in Society' research programme, under which the research in this paper has been funded (Award RES-160-25-0031 for the project **'**The World Wide Web of Science: Emerging Global Sources of Expertise'). We would also like to thank our interviewees.

References

Barabási, A. L. and Albert, R. (1999). 'Emergence of Scaling in Random Networks', *Science*, vol.296, pp.509-512.

Borgman, C. (2000). *From Gutenberg to the Global Information Infrastructure. Access to Information in the Networked World.* Cambridge MA: MIT Press.

Caldas, A. (2004). *The Structure of Electronic Scientific Communication: Electronic Networks, Research Collaboration and the Discovery of Digital Knowledge Bases*. Ph.D. thesis, University of Sussex, SPRU - Science and Technology Policy Research Unit.

Drori, G., Meyer, J., Ramirez, F. and Schofer, E. (2003). *Science in the Modern World Polity: Institutionalization and Globalization*. Stanford, Stanford University Press.

Dutton, W., Eisner Gillet, S. McKnight, L. and Peltu, M. (2003). 'Broadband Internet: The Power to Reconfigure Access'. *Oxford Internet Institute Forum Discussion Paper* no.1, available at www.oii.ox.ac.uk.

Edgerton, D. (1998). 'De l'innovation aux usages: Dix theses eclectique sur l'histoire des techniques' [From Innovation to Use: Ten Theses on the History of Technology], *Annales HSS*, vol. 4-5, pp. 815-37.

Frank, R. H. and Cook, P.J. (1995). *The Winner-Take-All Society: Why the Few at the Top Get So Much More Than the Rest of Us.* New York: Free Press.

Merton, R. (1988). 'The Matthew Effect in Science, II', *Isis*, vol. 79, pp. 606-623.

Nentwich, M. (2003). *Cyberscience: Research in the Age of the Internet*. Vienna, Austrian Academy of Sciences Press.

Pennock, D., Flake, G.W., Lawrence, S., Glover, E.J. and Giles, C.L. (2002). 'Winners don't take all: Characterizing the competition for links on the web', *Proceedings of the National Academy of Sciences*, vol.99, no.8, pp.5207-5211.

Wasserman, S. and Faust, K. (2004). *Social Network Analysis: Methods and Applications.* Cambridge: Cambridge University Press.