JOURNAL OF COMPUTER-MEDIATED COMMUNICATION



# Patterns of Information Search and Access on the World Wide Web: Democratizing Expertise or Creating New Hierarchies?*[1]

Alexandre Caldas

Affiliation: The Management Centre for the Electronic Government Network, Portugal

Ralph Schroeder

Oxford Internet Institute

Gustavo S. Mesch

University of Haifa, Israel

William H. Dutton

Oxford Internet Institute

*Will the World Wide Web and search engines foster access to more diverse sources of information, or have a centralizing influence through a 'winner-take-all' process? To address this question, we examined how search engines are used to access information about six global issues (climate change, poverty, HIV/AIDS, terrorism, trade reform, and Internet and society). The study used a combination of webmetric analyses and interviews with experts. From interviews we were able to explore how experts on these topics use search engines within their specialist fields. Using webmetric analysis, we were able to compare the results from a number of search engines and show how the top ranked sites are clustered as well as the distribution of their connectivity. Results suggest that the Web tends to reduce the significance of offline hierarchies in accessing information – thereby "democratizing" access to worldwide resources. It also seems, however, that centers of expertise progressively refine their specializations, gaining a 'winner-take-all' status within a narrower area. Some limitations of the winner-take-all thesis for access to research are discussed.*

doi:10.1111/j.1083-6101.2008.00419.x

## Introduction

The Web is rapidly developing into the largest worldwide sociotechnical system for the storage and dissemination of information on issues ranging from popular culture to science. On the Web, search engines have become an increasingly common tool for identifying useful sources of information and expertise. The algorithms

underlying search engines and the ways users employ these tools are thus likely to shape who gets access to what parts of the Web – with potentially far-reaching consequences for who knows what about any given issue as well as which sources become dominant for information about particular topics. How will search engines influence the ways in which researchers and others access information on the Web and what effect will this have on the relative prominence of alternative sources of expertise?

The growth of the Web over the last decade as a widely accessible medium of communication and a distributed system for knowledge sharing has been paralleled by extensive research on 'information retrieval' and Web search. The scale and rapid growth of the Web are challenging traditional methodologies for finding information of greatest relevance to the user. Nevertheless, search engines are becoming one interface of choice for access to this globally distributed information resource.

Technical features of the Web including global access make it possible for sources of expertise to be more decentralized since research centres around the world are accessible at any scientist's desktop. However, whether the Internet and the Web will in fact tend to democratize access to expertise is unclear. The purpose of this study is to investigate the extent that the use of alternative search technologies decentralizes access to scientific knowledge.

It has been argued, for example, that the Web exhibits power-law distributions of resources (see Barabasi and Albert 1999)[1] and differentiated patterns of connectivity and centrality[2]. This can be interpreted as having the implication that sources of information on the Web exhibit the characteristics of a 'winner-take-all' phenomenon (Frank and Cook 1995), an effect which has also been characterized as a 'cumulative advantage' or 'Matthew effect' in science communication (Merton, 1968). Others have argued that there is no such 'power law' or 'winner-take-all' effect (Pennock, Flake, Lawrence, Glover & Giles. 2002), and that the Web democratizes sources of information or there is an 'egalitarian effect of search engines' (Fortunato et al. 2006). It may also be that the winner-take-all or egalitarian effects oversimplify more complex patterns of access to information over the Web (see the discussion by Benkler, 2006: 241-61). By identifying the emerging forms of stratification on the Web, and whether these reinforce, democratize, or centralize sources of information, this paper contributes to an understanding of the uses of new tools for accessing expertise.

The next section of this article provides a brief review of research on Webmetrics, search engines and particularly on how users access information on the Web. This is followed by an outline of the hypotheses to be tested, and and overview of the methodology used for the analysis of empirical data. The final sections discuss the main results and outcomes emerging from the quantitative and qualitative examinations of search engine technology across six global topics, concluding with a summary of the main findings and suggestions for further research.

**Background: Webmetrics, Search Engines and their Uses**

The field of "Webmetrics" encompasses a wide range of theories, methods, and applications for measuring information on the Web (hence it can be considered also a type of "infometrics"). A taxonomy for Webmetrics would classify various measures and indicators of the Web into several different but complementary categories and include the following: graph properties (centrality and locality), significance (relevance and quality), similarity (content and link structure), search engines (effectiveness and comparison), usage, and other theories from information studies. In this study we focus on methods for analysing the Internet and the World Wide Web in terms of the structure of hyperlinks among Web resources.

Using the Web to search for information is becoming common. A recent study shows that most Internet users conduct searches using at least one search engine, and that on a given day, at least among American users, 56 percent of those online use search engines (Fallows, 2005). Asked differently, in Britain in 2005, about one-fifth (19%) of Internet users said they primarily rely on search engines, while another fifth (19%) primarily go to specific Web pages. Most (60%) rely to the same extent on both approaches (Dutton et al, 2005: 33). By 2007, well over half (57%) said they mainly use a search engine such as Google (Dutton and Helsper 2007: 66).

Reasons for using search engines vary a great deal. One study that took search terms from logs of three popular search engines and organized them into categories found that the top topics of search were (1) people and places, (2) commerce, travel and employment, (3) computers and Internet technologies and (4) health and sciences (Spink and Jansen, 2004). When using search engines, it is remarkable that about half of the users rely on a single search engine; those that use more than one tend to be more experienced, longer-term Internet users who have either developed more search skills over time, or began using a variety of search engines prior to Google's growing dominance.

Also, people tend to trust search engine results, with 68 percent of users evaluating search engines as reliable and unbiased, which might help to explain why users tend to stick to a single search tool (Fallows, 2005). Finally, it is important to recognize that although most Internet users rely on search engines, they also have other ways of retrieving information online, such as lists of their favourite sites, searching in portals, using links that have been recommended, or following links from other Web pages (Dutton et al, 2005; Fallows, 2005). But despite the rise of search engines as a means for obtaining information, there continue to be questions and challenges – both technical and particularly social - about how search engines work and with what effect on what users find online.

*Search Engine Technology*

Search engine technology is mainly concerned with providing a solution to two related functionalities: *precision (*or *relevance)* and *recall capacity*. The results of a search on the Internet should be as 'precise' as possible in order to be effective. The goal of search engines is that precise results should be available in a few seconds

for any given topic being searched. This goal is sometimes at odds with the number of items of information that a search can provide the user (the capacity for "recall" of the search engine), which can be defined as the number of relevant documents retrieved as a proportion of the total number of relevant documents available). Still, most assume that the *more* results returned the better. In short, search engine technology should simultaneously provide 'more' and 'very precise' items of information as an outcome of any search.

Taking into account the enormous size of the Web, and its dynamic nature, this is a very complex technological undertaking. Most search engines index documents in a database recovered from systematic crawls of the Web. The crawls explore the link structure of the Internet in order to jump from one document to related documents. Complementary content analysis provides a classification of the retrieved documents. However refined search engine technology has come to be, estimates suggest that 12 search engines taken together index less than 50% of the entire corpus of information on the Internet (Lawrence and Giles, 1999). As noted by Aquino and Mitchell (2001), search engine technology has used three different types of algorithms to build these indexed databases: the Naïve Bayes model, which focuses on topic-word frequencies; "maximum-entropy" algorithms, which focus on word combinations and how frequently they are associated; and, perhaps the most promising approach, the "cotraining" model, which studies the information on a Web page, as well as the linked pages, building an association of correlations.

These strategies for combining the content of Internet resources with the link structure of those resources are best suited for "entity extraction" – the ability to build databases from collections of specific entities. This brings us closer to the notion of "Web communities" and the self-organisation of the Internet. A Web community could be defined as a set of Web pages that link (in either direction) to more Web pages in the community than to pages outside of the community (Flake, Lawrence & Giles, 2000), making it useful to be able to identify such subsets of the large electronic network.

Several studies have empirically identified such communities based on a combination of several algorithms. One of the most common ones is the *Hubs* and *Authorities* resources in Internet Web space (HITS) algorithm (Kleinberg, 1998) which explores the link structure of the Internet, starting from a set of seed URLs and determining the HITS. *Hubs* are Internet resources that link to many authoritative pages in the topic, while *Authorities* are Internet resources that are linked by many Hubs. The methods (algorithms) for identifying Hubs and Authorities have been refined to overcome common problems. The PageRank algorithm implemented in the search engine Google, for example, uses some features of the initial HITS algorithm (Brin and Page, 1998; Huang, 2000; for search engines specialised in a specific topic area see e.g. Chau, Chen, Qin, Zhou, Qin et al., 2002). New models of Web search continue to be developed, in order to overcome problems, such as in optimising relevance and accuracy in search engines (Broder, 2002; Eastman & Jansen, 2003). A more difficult problem is related to the preservation of indexed databases

and the capacity of search engine to cope with Web "evolution" (Hellsten, Leydesdorff & Wouters 2006) – the fact that the Web is constantly evolving.

*Modelling the Structure of the Web*

Another set of efforts to provide more efficient and effective search engine technology focuses on modelling the structure and "networked" nature of the Web. The Web can be modelled as a large-scale and sparse graph (Kumar, Raghavan, Rajagopalan, Sivakumar, Tomkins & Upfal, 2000). A graph is a mathematical representation of a set of nodes (entities) and edges or arcs (links) among those nodes. Additional information (such as intensities or strengths of association) can be attributed to either nodes, or arcs, or both. More information can also be determined for subcomponents of the Web graph or for the whole graph, such as its size, connectivity, density, and clusterability.

A number of interesting properties have been identified through studies of the mathematical and statistical characteristics of Web graphs. These properties include the *distribution* of nodes in the graph, the *average path distance* among nodes within the graph, and the *clustering coefficient* of each node and the graph as a whole. These features are important to our analysis of the distribution and clustering around our selected topics.

Nodes in the Web graph possess different and significantly skewed *out-degree* and *in-degree characteristics* (the number of links *originating in* a certain node, and the number of hyperlinks *directed to* a certain node). Power-law distributions and variations of them have been discovered to characterise the Web graph and subcomponents of it (Kleinberg, 1999) whereby a small number of links are linked to very large number of other links while a very large number of links are linked to a very small number of other links.

This property can be related to a second property, the 'scale-free' nature of the Web; that is, the notion that patterns that can be found to characterise small-scale parts of the Web can also be found to characterise it at larger scales. In analyzing the structure of links for our six topics, we found that clustering, power laws and the scale-free nature of the Web combine to allow us to identify *cliquishness* in how the Web has become organized around the various science topics.

*Information Seeking, Communication, and Collaboration on the Web*

The ways in which users access and share information within electronic networks and in particular via the Web, as well as how communication patterns emerge within particular communities of practice, might impact the efficiency and efficacy of search engine technology. There are a number of approaches to investigating how the use of the Web is changing how people access information. Perhaps the broadest is the notion of 'reconfiguring access' (for example, Dutton 2005), which encompasses how new information and communication technologies reshape not only how people access information, people, services, and technologies, but also the outcomes of these processes. How ICTs reconfigure access depends on rational strategies in

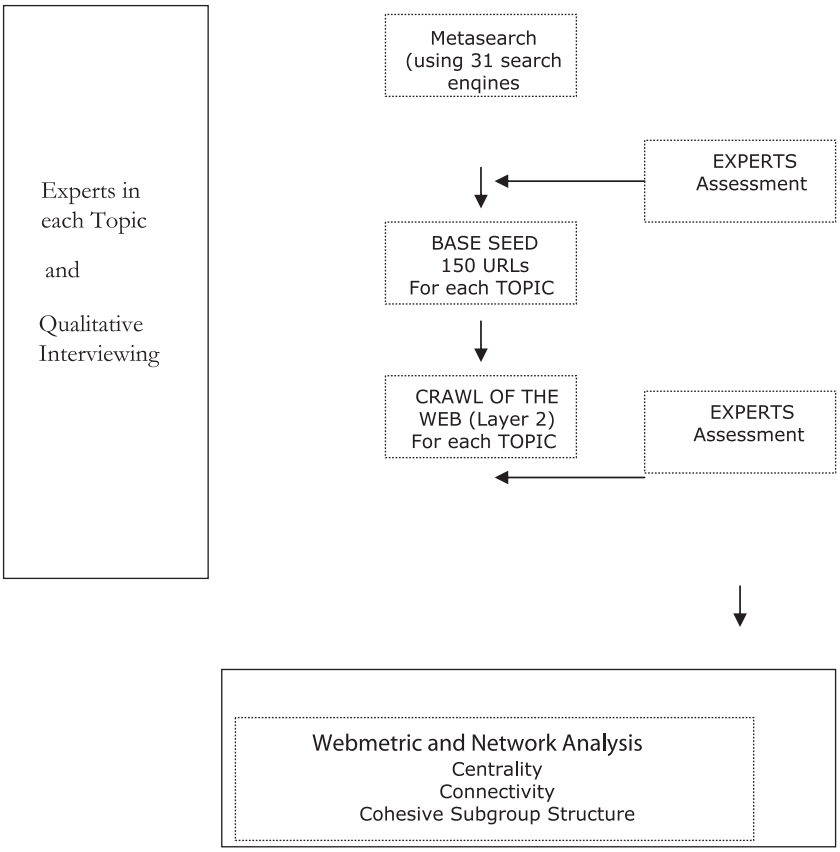


**Figure 1** Combination of Qualitative and Webmetric analysis with a Metasearch strategy.

seeking information and on less conscious choices such as people's information habits – but also on how technologies are shaped by existing socioeconomic patterns. In this study, we are particularly interested in how researchers use the Web and how the Web has become an indispensable tool for keeping abreast in their fields.

The exchange of information between scientific researchers, or scholarly communication, has been extensively analyzed even before the arrival of the Internet and the Web. This raises the question of what has changed, or how the online world of the Web corresponds with the offline relations between scientists? Offline networks of scientists have been discussed extensively in the sociology of science (for a review, see Hess 1997). The online world, with its links and techniques for finding different places, has a structure of its own, and so shapes - even if it does not determine - access to different Web locations. Traditionally, for offline publications, bibliometric analysis has been one method to identify networks of scientists working together and to identify the links between their research outputs. Now, with online tools, this method may be extended to online resources and to Web links in particular (see Borgmann and Furner, 2002 for an overview). If search engines are used to access

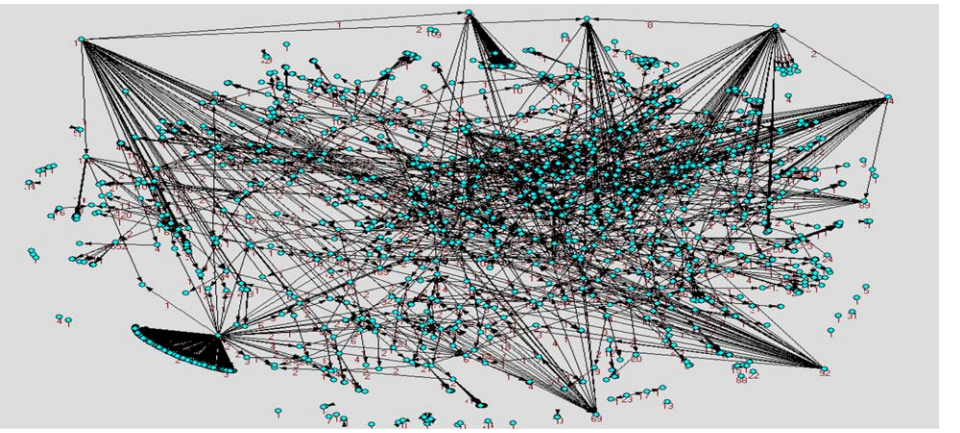

**Figure 2** Web space graph of 150 URLs in Climate Change.

information about research, however, they may serve as new 'gatekeepers,' reconfiguring access to the world of online research. Apart from examining the networks of online links then, we will need to examine how researchers use search engines to access information and knowledge.

In addition, studies have examined different aspects of the uses of online tools by researchers. Matzat (2004) analyzed the effect of Internet discussion groups on academic communication. He found that these discussion groups do not equalize the relations between peripheral and well-integrated researchers, but that they are useful for the creation of new social contacts. One of his findings of particular relevance to this study is that Internet discussion groups do not counteract the Matthew effect (Merton, 1968) whereby well-established researchers are cumulatively more advantaged than the less established, supporting a winner-take-all hypothesis (2004: 246).

Walsh, Kucker, Maloney, and Gabbay (2000) compared the use of computer-mediated communication (CMC) in four different scientific fields and for different research related activities and found that the use of e-mail enhanced scientific collaboration and productivity. Other studies have taken a more holistic view, examining research and information seeking practices across a range of electronic resources and comparing how different disciplines make use of these resources. Fry finds (forthcoming, see also Fry and Talja 2004), for example, not only that there is great variation between scientific disciplines, but also within disciplines and between the use of different electronic tools and resources. Fry's research meshes well with Nentwich's analysis of different scientific disciplines and their degrees of 'cyberness,' or the degree to which they make use of ICTs and particularly the Internet, which varies between disciplines (2003:37). And against expectations, as in Fry's research, the main divide is not between natural sciences with a high degree of 'cyberness' and the humanities with a low one, but more varied within these two broad categories.

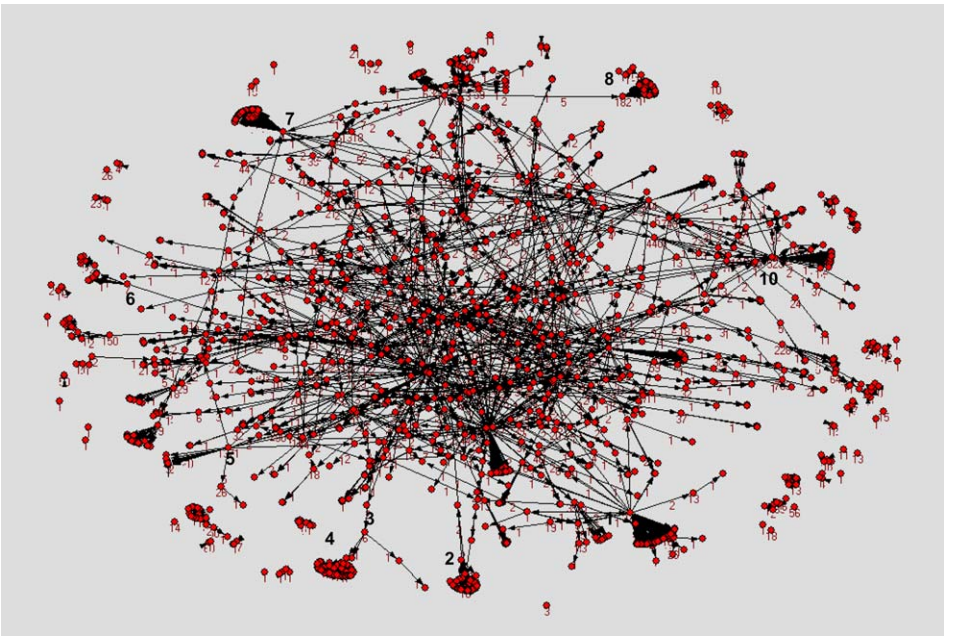

**Figure 3** Web space graph of 150 URLs in Internet and Society.
**Legend:**
1 – http://www.ercim.org (European Research Consortium for Informatics and Mathematics)
2 – http://www.apa.org (American Psychological Association)
3 – http://www.aoir.org (Association of Internet Researchers)
4 – http://www.iiiib.org (International Institute for Internet Industry Benchmarking)
5 – http://nsr.mij.mrs.org (Internet Journal of Nitride Semiconductor Research)
6 – http://www.ecommons.net (Electronic Commons, Canada)
7 – http://lcweb.loc.gov (Library of Congress, US)
8 – http://www.senate.gov (US Senate)
9 – http://www.aciri.org (Centre for Internet Research, Berkeley)
10 – http://www.gatech.edu (Georgia Institute of Technology)

Finally, it has been argued that Webmetric research on links between Websites provides a powerful tool for analyzing scientific communication (Park and Thelwall, 2005), though Beaulieu (2005) argues that a qualitative approach which analyses texts and puts them into context is necessary—as opposed to the quantitative analysis of networks of links proposed by Park and Thelwall (2005). In our study, we have relied mainly on techniques derived from research on search engine tools and Webmetric analysis, though we complement this with qualitative interviews (see Fry, Virkar, and Schroeder 2008) and information from a limited number of experts.

Independent of how offline social structures are reproduced by the Web, it is likely that search engine technology will enhance global access to knowledge resources. Different uses of search engine technology will also lead to new and innovative

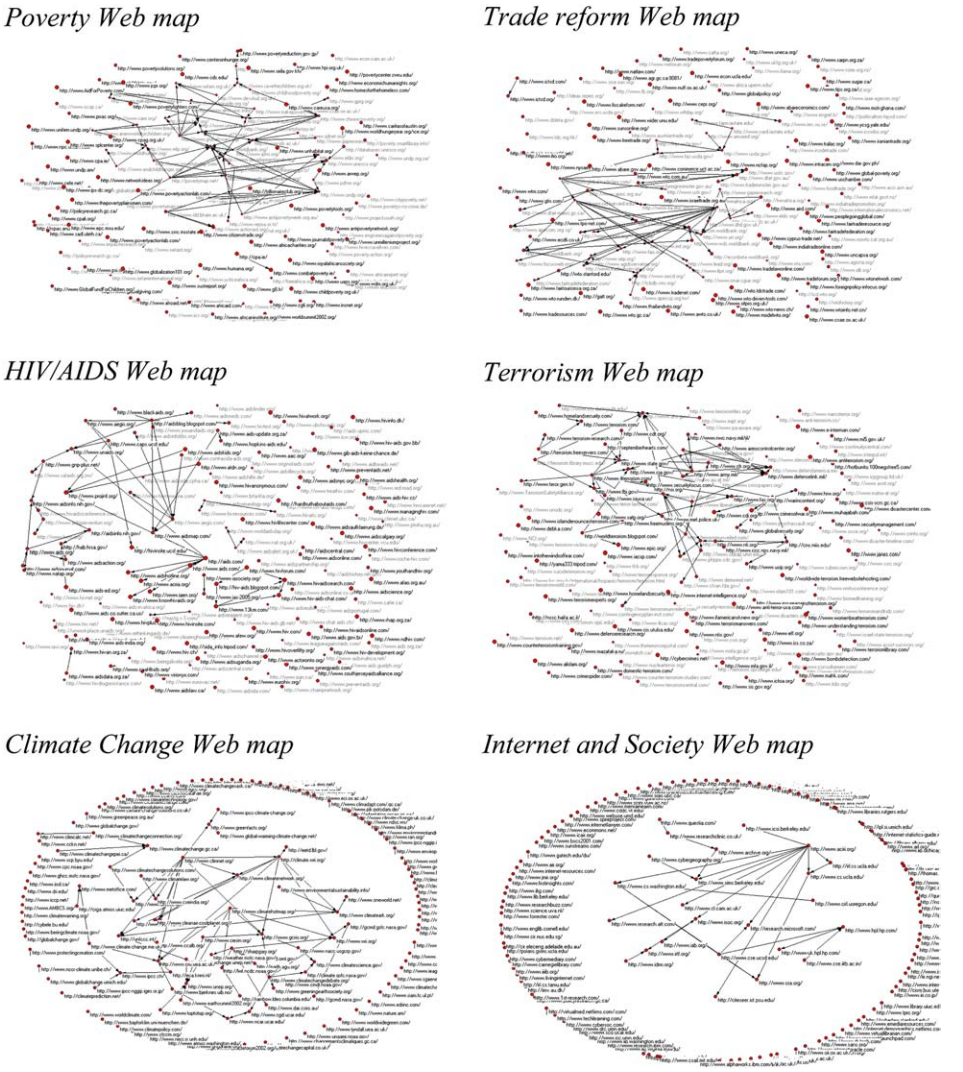


**Figure 4** Networks representing the interlinkages among 150 selected nodes in each of the six global topics.

**Legend:**

From top left to right bottom: Poverty, Trade Reform, HIV/AIDS, Terrorism, Climate Change and Internet and Society.

patterns of information creation, distribution and use. This is particularly true if new, more decentralized, or peer-to-peer organisational forms of knowledge sharing are created in the future and if there are further advances in search engines.

Despite the growing body of research on how users' access knowledge on the Web, on the technical and social challenges of search engine technology, and how the

structure of the Web affects scientific communication, we still know little about how the structure of knowledge represented on the Web effects the ability of researchers to access and find information. In particular, how search engines work and the power laws that govern results could exercise a powerful effect on what is found. As there has been little research on this topic, we formulated a number of tentative hypotheses to empirically examine in this study.

## Hypotheses

As discussed above, a number of factors shape how information will be found on the Web: the overall structure of the Web, the mechanisms embedded in search engine technology, and how researchers seek information in the online world about global topics. The combination of the structure of the Web, distribution of content on the Web and Web usage patterns is likely to have a significant impact not only one how people access information on the Web, such as through search engines, but also the outcomes of their search. To investigate how search engines will configure access to information, we therefore put forward two complementary hypotheses.

> H1: Because of differences in the design of the technical features of search engines, in terms of their capability for recalling results across the various search topics, different search engines will yield significantly different results.

Support for this hypothesis might lead researchers to consider, for example, using different search engines and different search strategies depending on the topic and the purpose of the search, though another approach could be to use a combination of search engines along with other search strategies.

> H2: It is possible to identify regularities and patterns in the distribution of Web resources for various search topics, particularly in terms of centrality, connectivity and subgroup structure.

For example, network structures of the Web or its "Web communities" around a particular topic might exhibit 'Power law' features – that is, a hierarchical structure in the distribution of links, but they might also reveal a "fractal structure," a replication of certain network patterns on different scales. Support for this hypothesis should focus more attention on how the search engine technology reproduces certain structures of links and how these structures relate to content and information-seeking behaviour.

In addition to these two central hypotheses, this study also investigated a number of questions of a more exploratory nature by means of analyzing our qualitative data: How do Web searches produce different results depending on different approaches to a topic? And – how do certain characteristics of the researchers, such as their geographical location or the combination of online resources they use, affect their search results?

## Methods

We examined these hypotheses through a combination of methods.

Initially, we compared searches using six different search engines (Google, Yahoo, MSNSearch, AskJeeves, Gigablast, and ScholarGoogle) for a set of three keywords for each topic. Search engines were selected by authors according to their importance, and in order to guarantee in the study different types of search engine (generic: e.g. Google and MSNSearch, and AskJeeves as opposed to directory: e.g. Yahoo; and specific: e.g. Gigablast and ScholarGoogle). A second and complementary method was to use a "metasearch engine" combining 31 search engines[3] for an extended set of keywords in each topic ("extended search"). Following from this metasearch strategy, a "structurally embedded" analysis (analysis of the link structure) was conducted based upon the Webmetric information resulting from network analysis of Web linkages on the Web in the six global topics.

This quantitative method was combined with qualitative expert assessment and validation of the results on two of the research topics. Experts were invited to participate for reasons of proximity (to be based in Oxford and expertise in their fields) as personal engagement was considered to be a fundamental criterion at this stage of the study. This provided a means of checking Webmetric results against expert assessments. The two topics on which this expert assessment was piloted ('Internet and society' and 'Climate change') allowed researchers to assess the lists of URLs and gave us a good indication of the validity of the results of searches and of the link structure that was discovered. The following diagram provides an overview of the methodology. This combination of quantitative and qualitative provided a robust way of testing our hypotheses and validating them for two of the topics. This combination of methods will provide a good basis for future research.

### Data Analysis

This section presents the results for the empirical analyses. First, we show the outcomes for the search engine analyses. Secondly, we discuss the Webmetrics of the structure of Web links. And finally we interpret our pilot study on qualitative assessment by our small group of experts.

#### *The Heterogeneity of Search engine Technology*

The results from querying different search engines for our six topics confirm hypothesis 1 with respect to the heterogeneity of the results of search engines. Similar patterns and results are consistent for the six topics.

Table 1 shows that although there are some similarities in the overall magnitudes of the results for the different topics, there are also some significant differences between the search engines. First, the capacity of search engines to recall results differs significantly. For example, the potential total number of results on any query coming from each search engine is considerably different from using Google or AskJeeves or MSNSearch. In this respect, we find support for the first hypothesis that search engines have different "sizes" and "scale" differently.

**Table 1** Heterogeneity of Search Engine Results

Search-Engines Results for Six Global Topics

| | Google | Yahoo | MSNSearch | AskJeeves | Gigablast | Scholar.Google |
|---|---|---|---|---|---|---|
| **Terrorism** | 51,000,000 | 103,000,000 | 11,030,218 | 7,185,000 | 73,936 | 86 |
| **HIV/AIDS** | 26,200,000 | 44,600,000 | 6,754,236 | 7,099,000 | 196,608 | 156,000 |
| **Climate Change** | 34,500,000 | 37,400,000 | 6,566,133 | 6,363,000 | 72,178 | 227,000 |
| **Internet and Society** | 1,230,000 | 4,430,000 | 763,500 | 313,100 | 57,757 | 8,710 |
| **Trade Reform** | 1,040,000 | 2,220,000 | 496,267 | 183,700 | 56,850 | 20,700 |
| **Poverty** | 142,000 | 348,000 | 80,337 | 35,600 | 15,742 | 113 |
| Self-Reported Index Size | **8,168,684,336** | * | **5,000,000,000** | * | **2,024,193,536** | * |

* Data not available

Search engines were queried in the period 8 - 10 August 2005, using the following Keywords for each of the six global topics

**URLs:** www.google.com; www.yahoo.com; search.msn.com; www.askjeeves.com; www.gigablast.com; scholar.google.com

**Keywords:**

Internet and Society: "Internet and society" OR "Internet research" OR "Internet studies"

Climate Change: "Climate change" OR "Global Warming" or "Ozone Depletion"

Terrorism: "Terrorism" OR "Terrorist organisation" OR "Terrorist network"

Trade Reform: "Trade reform" OR "Trade liberalisation" OR "Trade and development"

Poverty: "Poverty research" OR "Poverty statistics" OR "Poverty and globalisation"

HIV/AIDS: "HIV/AIDS" OR "HIV infection" OR "HIV prevention"

Secondly, there are considerable differences in the quality and "genre" of results across the search engines. In this regard, Google, MSNSearch, and AskJeeves form a distinct cluster, tailored for "general search" without much focus but with very-large scale capabilities. A second and distinct group is formed only by Yahoo which represents a different kind of search engine characterised by being a "directory" organised by themes and topics, but providing "directory" oriented search engine functionalities. A third group is formed by Gigablast, a specialised search engine more focused on scientific and technical information, providing wider search functionalities but focused on connectivity among documents on the Web – 'document-based' search. Finally Google Scholar constitutes a completely distinct search engine, only focused on indexing scientific and technical literature.

Thirdly, we can see that there are important differences across the search engines in terms of the "content" indexed in the various crawlers and indexing technologies. A detailed analysis of the outcomes of the 30 first results for each topic and across the search engines demonstrated this "diversity" in content (see Table 2 below for a comparison of results).

*The Structure and Hyperlinked Nature of Knowledge on the Web: Webmetrics and Metasearch Strategy*

Against this background of the overall differences between search engine results, we can turn to whether these results are linked to an underlying structure of links on the

**Table 2** Heterogeneity of Content and Results Across the Six Topic Areas

| Topic | Common URL Results (between at least two Search-Engines) | | | | |
|---|---|---|---|---|---|
| | Google | Yahoo | MSNSearch | AskJeeves | Gigablast |
| **Terrorism** | | | | | |
| www.terrorism.com | X | | X | X | |
| www.terrorism.net | X | | X | X | |
| www.bt.cdc.gov | X | | | X | |
| www.cia.gov/terrorism | X | | X | | |
| www.whitehouse.gov | X | | | X | |
| en.wikipedia.org/wiki/Terrorism | X | | X | | |
| www.mipt.org | X | | X | X | |
| www.ict.org.il | X | | | X | |
| | | | | | |
| **HIV/AIDS** | | | | | |
| www.unaids.org | X | | | X | |
| www.aidsinfo.nih.gov | X | X | | | |
| www.thebody.com [1] | X | X | | X | X |
| www.cdc.gov/hiv/dhap.htm [2] | X | X | X | X | |
| hivinsite.ucsf.edu | X | X | X | X | |
| www.un.org/ga/aids/coverage [3] | X | X | | | |
| www.aegis.com [4] | X | | | X | X |
| www.hopkins-aids.edu/ [5] | X | | X | | |
| | | | | | |
| **Climate Change** | | | | | |
| www.ipcc.ch | X | | | X | X |
| www.globalwarming.org | X | | X | X | |
| www.climatehotmap.org | X | | | X | |
| www.climateark.org | X | | X | X | X |
| www.pewclimate.org | X | | X | X | |
| www.epa.gov/globalwarming | X | | X | X | |
| en.wikipedia.org/wiki/Global_warming | X | X | | | |
| unfccc.int | X | | | X | |
| | | | | | |
| **Internet and Society** | | | | | |
| www.jmir.org | X | X | X | X | X |
| cyber.law.harvard.edu | X | | X | X | |
| www.irtf.org | X | | | X | |
| cyberlaw.stanford.edu | X | | X | X | |
| www.aoir.org [6] | X | | | X | X |
| www.internetstudies.org | X | | X | | |
| www.isoc.org | | | X | X | |
| www.isc.umn.edu | X | | X | X | X |
| | | | | | |
| **Trade Reform** | | | | | |
| www.unctad.org | X | | X | X | |
| www.tda.gov | X | | X | X | |
| www.itd.org | X | | X | X | |
| www.tln.org.nz | X | | X | X | |
| www.un.org [7] | X | | | X | |
| www.wto.org [8] | X | X | | X | |
| | | | | | |
| **Poverty** | | | | | |
| www.jcpr.org | X | X | X | X | X |
| www.chronicpoverty.org | X | | X | X | |
| www.gapresearch.org | X | | X | X | |
| www.gprg.org | X | | | X | |
| www.sussex.ac.uk/Units/PRU | X | | | X | |
| www.census.gov/hhes | X | | X | X | |

The 30 first URL resources were selected for each of the six topics across the various search-engines

**Notes:**

1 URL match is partial as in Gigablast reference is to more detailed resource (www.thebody.com/cdc/factadol.html)

2 URL match is partial as in MSNSearch reference is to different resource (www.cdc.gov/hiv/pubs/brochure/Oi_toxo.htm)

3 URL match is partial as in Yahoo reference is to more detailed resource (www.un.org/ga/aids/coverage/FinalDeclarationHIVAIDS.html)

4 URL match is partial as in Gigablast reference is to more detailed resource (http://www.aegis.com/pubs/woalive/1993/WO1993-0902.html)

5 URL domain is referenced in both search-engines, but not the two different and detailed URL resources

6 Gigablast provides 3 results for the domain www.aoir.org - 2 with detailed resources (e.g. www.aoir.org/2002/program/feminist_futures.html)

7 URL domain is referenced in both search-engines, but not the two different and detailed URL resources

8 URL domain is referenced in the three search-engines, but not the distinct URL resources

Web. To do so, we present the results from a "metasearch," combined with Webmetric link analysis for our six topics. The results confirm interesting regularities across the whole set of topic areas, with regard to centrality, connectivity and subgroup structure of these "Web networks," and thus provide support for our second hypothesis that "social networks" are embedded in the distribution of Web resources, with important implications for search engine functionality and access to knowledge on the Web.

*Fragmented – fractal structure – of Web domains*

The experts we consulted provided us with a list of keywords for each of the six topics. On this basis, we randomly drew, from an initial list of 3,594 Web links collected from 31 search engines, a sub-sample of 1,156 Web references for "Climate Change." From this reduced subsample we were able to select 150 'good' URLs.[4] A "Web space graph" for these 150 URLs was then calculated by examining the *inlinks* and *outlinks* on the Web. A total of 3,489 distinct nodes was calculated. A graphical representation of the structure of this Web community is provided in Figure 2.

When we mapped the other five topics using the same procedure ('Internet and society,' 'poverty,' 'HIV/AIDS,' 'terrorism,' and 'trade reform'), similar patterns of centrality and connectivity emerged. These patterns testify to a 'fractural structure' of online networks, a structure which is replicated across the six areas and with similar patterns occurring in each area whereby the nodes in the network form hubs with (several) spokes connecting to these 'hub' nodes. The overall network then appears to form a 'fractal' whereby each 'subgroup' is interlinked by a reduced number of connections to other more distant groups. Even if there are differences across the six topics, the same 'abstract' structure of a fractal can be consistently identified across these Web spaces. Figure 3 provides an additional graphical representation of a Web graph for "Internet and Society". In a similar way to "Climate change", it is clear that there is a similar "fractal like" or clustered structure to these links.

The qualitative assessment of the relevance of these results to the field of study, based upon interviews with our expert panel, was not conclusive, but it provided some indication of the usefulness of Web mapping for identifying key entities on the Web. For example, many sources of information could be missing due to our sampling of Web sites. At the same time, some of these URLs identified were considered to be important sources of information about "Internet and Society," such as the Association of Internet Researchers (AoIR), The Georgia Institute of Technology, the Centre for Internet Research at Berkeley, the Library of Congress in the US, and the American Psychological Association in the subfield of communication studies. Overall then, it is possible to use Webmetric analysis as a first stage, but this will need to be complemented by other methods (for example, wider surveys of experts to identify seed sites) to achieve a robust mapping of the "Internet and Society" Web space. The same applies to "Climate Change."

The analysis of interlinkages on the Web also provided support for the hypothesis that there are some similar patterns within these electronic networks in terms of the centrality of certain resources, connectivity and subgroup structure. Table 3 below

**Table 3** Regularities in Web Networks: Average Distance, Connectivity, and Clustering for 'Climate Change' and 'Internet and Society'

| | 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. | | |

provides a comparison between "Climate change" and "Internet and Society" for relatively similar size Web networks. Comparing, for example, the indicators for average distance among reachable pairs of nodes, this is relatively small (2.2 and 2.3 respectively), which suggests that despite the significant size of these Web networks - 3,489 nodes and 20,839 arcs in "Climate change", and 3,815 nodes and 31,736 arcs in "Internet and Society" - there is nevertheless a short distance interconnecting any two different nodes on the Web.

Yet despite the differences in size and evolution of these systems, there was a significant process of "clustering" within these electronic networks and we can indeed identify subgroups of tightly knit and highly clustered nodes. The factors which explain these relatively high "clustering" coefficients can probably be better explained by offline characteristics than by the online link structure.

Finally, it is noteworthy that these networks exhibit low density and a 'sparse' nature. In any of the six Web networks there are a large number of nodes that are loosely connected or even isolated, and a much smaller number of nodes that are centrally connected and in a more highly interconnected 'core'. Overall, links on these six topics formed a low-density connected network on the Web. In short, this 'fragmented" nature of these Web networks could point to a 'democratization' (low density) and 'reinforcement' (clique) effects rather than to a 'power law/winner-take-all' effect. However, without additional results for other networks, more longitudinal study, and triangulation of results by other methods, these are merely pointers that will need to be investigated further.

*Link Structure of the Web*

Figure 4 shows the interlinkages only among 150 selected "core nodes" across the six topics. As Figure 4 illustrates, similar structures were found for all the six topics (although these small images are only indicative; the data presented in table 3, and

larger images would give, a better indication). This illuminates the low-density connected nature of these electronic networks, as well as the "core-periphery" type of structure emerging in these Web networks. Combined with what we know about the structure of the Web, as discussed above, these Web maps reinforce the low-density and highly clustered nature of networks on the Web. Perhaps more importantly, the "social networks" that apparently explain the structured nature of these networks seem to be a consistent feature of the way entities and resources appear to be interlinked on the World Wide Web. It should be useful for search engine technology to make such a major characteristic of the Web apparent to users.

*Reproducing offline characteristics*

Combining the validation of our expert groups on "Internet and Society" and "Climate Change" (see below) with our Webmetric findings about these online structures, suggests that "offline social structures" (the high status and visibility of certain institutions) are key drivers for the distribution of online resources and Web access to information. For example, we found, in our interviews with expert groups on "Climate Change" and "Internet and Society," that some fundamental characteristics of how users traditionally access information offline, as well as essential variables such as geography, locality, and influence on "policy," heavily influence the distribution of Web resources.

For example, several of the researchers from the expert group on "Climate change" pointed to the importance of geography and locality for how they access resources in their field and how they are distributed, also because of how the importance of climate change varies for different places of the world. Search engines are 'neutral' in relation to the landscape of research in the sense that sites are singled out for reasons of their political geography. For 'climate change,' however, several of our UK interviewees said that they accessed more European (including UK) sites related to their research, and the reason they gave is political: Since the US administration has taken a stance against urgent action on climate change (as with its stance against the ratification of the Kyoto treaty) whereas the European (and worldwide) position has been to take a more proactive role, our UK researchers thought that this oriented them more towards European research. This suggests a more general political dimension to search strategies within the sciences that is worthy of further exploration. It may be that the relatively apolitical nature of search algorithms will improve Web-based search over traditional personal selections.

In other words, "academic content" needs to be supplemented with a view on the societal interest in information on the Web. Both the "Internet and Society" and the "Climate Change" expert groups mentioned that there was bound to be a difference between "academic networks" as opposed to Web resources that were more oriented to a wider audience interested in these social issues.

*A Pilot Qualitative Assessment on the Use of Search Engines*

We focus here on two of the six topics 'Internet and Society' and 'Climate Change' for reasons of space (interviews for the other topics are discussed in Fry, Virkar, and

Schroeder 2008). These two areas represent two quite different domains in the social and natural sciences. We summarize material from two sources: the first are interviews (both individual and in small groups) with eight 'Internet and society' researchers and six 'climate change' researchers, and the second are lists which these interviewees compiled of the sites they use in their areas of expertise and those that they rank as the most important sites.

We asked our interviewees to provide us with information on how they search for information in the research projects they are currently conducting. In some cases, the search started not on the Web but by first using other academic online resources, such as searching the university library database, scientific databases, and citation indexes. After relying on theses sources, however, all our researchers used search engines. One of the researchers said "I think that everybody in my field uses a variety of sources to gather literature and data on the research, but eventually all of us end up using a search engine, to make sure that we do not miss unpublished manuscripts or manuscripts that did not came up in other sources."

The other main point arising from our interviews is that these researchers all primarily use Google (and some occasionally Google Scholar). They use other search engines rarely or only secondarily for special purposes. Obviously it is not possible to generalize from such a small sample, but Google was the dominant search engine by far among the researchers interviewed (Google is also the most frequently used search engine generally, with a share of approximately one-third of unique users, see www.searchenginewatch.com, but for caution about these figures, see Hargittai, 2004).

Convenience, the ability to have the information at one's desk and be able to download the material, is apparently a key factor for preferring search engines to other modes of information seeking. For example, one of our interviewees said that "the first material that I read is [sic] the one that is online. When a paper or article is not online, a copy must be made at the library, so I put them on the side and the reading is postponed …and sometimes I do not read [the library material] at all".

As with other uses of the Internet it seems that researchers become used to a particular search engine after using it over a period of time. Our small group of researchers also told us that although they were aware that there would be differences in the results from different search engines, they regarded these differences as marginal and not important enough for them to pursue. Finally, they indicated that they trusted the results provided by search engines.

Against our expectations they used search engines in more idiosyncratic ways, with no discernable common features within or between the two domains ('Internet and Society' and 'Climate Change'). Nor did any two researchers exhibit the same search behaviour. For example, when exploring the different strategies used, some researchers indicated that they are more person-oriented in their searches (they look up names of individual researchers) and others more institution- and topic-oriented

searchers, but this does not depend on the particular topic or the area within the research. In terms of the keywords that people used, we expected 'climate change' researchers to use the names of institutions related to the topic more often compared to 'Internet and society' researchers who would be more likely to use the names of individual researchers. Yet in fact, different interviewees from both groups in some cases searched primarily by using names of institutions, or by using the names of individual researchers, and in other cases they used keywords related to the research topic itself. And although some used one type of keyword search in preference to the others, most researchers used all three - institution names, researcher names, and topic-specific keywords.

All of our interviewees used the most popular English-language search engine, Google, to identify sources that are not available by means of other sources, for example to check material that is not available via online libraries and databases. They also used Google *in combination with* other types of search, for example with Google scholar, with metacrawlers (aggregated search engine search machines), or in the case of one 'Internet and Society' researcher in combination with a search engine (dogpile.co.uk) that is specifically aimed at government sites and official publications in the UK (his specialist area). Or again, in the case of several of our 'climate change' interviewees, they used Google to look for contact information on institutional sites, pointers to databases, or to seek archived newspaper information. Finally, most of our interviewees used Google at all stages of their search for information, but one 'Internet and society' researcher said that she mainly used Google once other online sources (online articles and databases) have been consulted and mainly to follow up and look for authors' Web pages.

Thus, all researchers say that they use a combination of search engines and other electronic sources, going back and forth frequently between them. Against this background, we now present a list of the top 30 sites (Table 4) from among the top 150 identified by our Webmetric analyses that our climate change researchers said they knew about and that they referred to at least occasionally.

The results indicated that some "key" institutions or entities are regularly monitored through the Web. Next we asked our researchers to select, from the list of 150 URLs, those institutions which, although not used by respondents, were likely to be recognized as key sources of information, based on their importance in the (offline) world. The 15 institutions they identified, together with their URLs, are listed in Table 5.

These results indicated that there were a significant number of online sources that - even if they were not regularly used by our expert panel, apparently were likely to have some significance as they were easily recognised as offline organisations (e.g. see table below). This connection between the offline significance of some institutions and their online presence thus reinforces the notion (expressed in our second hypothesis about denser 'subgroups' of highly connected sites) that there are likely to be some groups of highly interlinked sits that will be highly visible via search engines.

Furthermore, we can see that there is a considerable "match" between the URL sources validated qualitatively by our expert panel of interviewees and the

**Table 4** Selected 30 URLS validated by "Climate Change" Researchers

| URL | Domain | Title | Know about | Usage |
|---|---|---|---|---|
| http://www.ipcc.ch/ | .ch | Intergovernmental Panel on Climate Change | 1 | 2 |
| http://www.pewclimate.org/ | .org | Global Warming: The Pew Center on Global Climate Change | 1 | 2 |
| http://www.metoffice.com/ | .com | Met Office homepage - The home of the Met Office on the internet | | 2 |
| http://www.unfccc.de/ | .de | United Nations Framework Convention on Climate Change | 1 | 2 |
| http://www.wri.org/ | .org | World Resources Institute | 1 | |
| http://www.ipcc-climate-change.org/ | .org | Climate Change : summary of the IPCC report | 1 | 2 |
| http://www.eci.ox.ac.uk/ | .uk | Environmental Change Institute | | 2 |
| http://www.pik-potsdam.de/ | .de | Potsdam Institute for Climate Impact Research | | 2 |
| http://www.earthsummit2002.org/ | .org | Earth Summit 2002, Johannesburg Summit, Rio+10, Johannesburg 2002, World Su | 1 | 2 |
| http://www.cru.uea.ac.uk/ | .uk | Climatic Research Unit | 1 | 2 |
| http://www.tyndall.uea.ac.uk/ | .uk | Tyndall Centre for Climate Change Research | 1 | 2 |
| http://climateprediction.net/ | .net | ClimatePrediction.Net gateway | | 2 |
| http://europa.eu.int/comm/environment/climat/ho | .int | European Commission, Climate Change | 1 | 2 |
| http://europa.eu.int/comm/environment/climat/es | .int | European Climate Change Programme | 1 | 2 |
| http://www.nbu.ac.uk/iccuk/ | .uk | Indicators of Climate Change in the UK | 1 | 2 |
| http://www.metoffice.gov.uk/research/hadleycen | .uk | Hadley Centre | 1 | 2 |
| http://www.ukcip.org.uk/ | .uk | UK Climate Impacts Programme | 1 | 2 |
| http://www.theclimategroup.org/ | .org | The Climate Group | 1 | 2 |
| http://www.scidev.net/ | .net | Science and Development Network | 1 | 2 |
| http://www.iisd.org/ | .org | International Institute for Sustainable Development | 1 | 2 |
| http://yosemite.epa.gov/oar/globalwarming.nsf/c | .gov | US EPA: Global Warming | | 1 |
| http://www.ippr.org.uk/ | .uk | Institute for Public Policy Research | 1 | 2 |
| http://www.iea.org/ | .org | International Energy Agency | 1 | 2 |
| http://www.undp.org/seed/eap/activities/wea/ | .org | World Energy Assessment Report | 1 | 2 |
| http://www.worldbank.org | .org | World Bank | 1 | 2 |
| http:www.doe.gov | .gov | US Department of Energy | 1 | 2 |
| http://www.parliament.uk/parliamentary_committ | .uk | House of Parliament: Science and Technology Select Committee | 1 | 2 |
| http://www.sei.se/ | .se | Stockholm Environment Institute | 1 | 3 |
| http://www.energyinst.org.uk/ | .uk | Energy Institute | 1 | 3 |
| http://www.teriin.org | .org | The Energy Research Institute, India | 1 | 2 |

**Know About?**
1 - Yes know about web site
0 - No
**Intensity of Use**
1 - Never
2 - Occasionally (every few months)
3 - Frequently (weekly or daily)

quantitative results coming from the automatic search of Google and the other search engines (compare table 4 and 5 with table 2).

Since our researchers overwhelmingly used Google (self-reporting from the interviews), we cannot compare the sites they made use of with the data from search

**Table 5** Selected 15 URLs on "Climate Change" - Identification Based on "Offline" Existence of Organisation

| URL | Domain | Title |
|---|---|---|
| http://www.greenpeace.org.uk/ | .uk | Environmental Issues \| GM Food \| Nuclear Power \| GREENPEACE UK |
| http://www.protectingcreation.org/ | .org | Interfaith Climate Change Network |
| http://www.clickforcleanair.org/ | .org | David Suzuki Foundation: Climate Change |
| http://earth.agu.org/ | .org | AGU. American Geophysical Union. Earth - Oceans - Atmosphere - Space - Planets |
| http://www.cmdl.noaa.gov/ | .gov | Climate Monitoring and Diagnostics Laboratory |
| http://climatechange.unep.net/ | .net | Climate change: UNEP.Net, the Environment Network |
| http://www.dar.csiro.au/ | .au | CSIRO Atmospheric Research |
| http://www.aoml.noaa.gov/ | .gov | DOC/NOAA/OAR/AOML: Atlantic Oceanographic and Meteorological Laboratory |
| http://crga.atmos.uiuc.edu/ | .edu | Climate Research Group Home Page |
| http://gcmd.nasa.gov/ | .gov | Earth Science data and services directory: Global Change Master Directory Web Site |
| http://www.protectingcreation.com/ | .com | Interfaith Climate Change Network |
| http://lwf.ncdc.noaa.gov/ | .gov | NCDC: * National Climatic Data Center (NCDC) * |
| http://www.ncar.ucar.edu/ | .edu | National Center for Atmospheric Research (NCAR) |
| http://www.bayforklim.uni-muenchen.de/ | .de | Bayerischer Klimaforschungsverbund BayFORKLIM (climate changes in Bavaria, Germany, ...) |
| http://www.cpc.noaa.gov/ | .gov | Climate Prediction Center |

engines other than Google. However, the results from Google and from other search engines differed. This difference is important on several dimensions, such as the "recall" of the different search engines and different results they provide.

## Conclusions and Further Research

The results discussed in this paper indicate the need for a more nuanced view of the efficacy of search engines and how they provide access to knowledge on the Web. On the one hand, search engine technology has evolved significantly in recent years and has become one of the most effective and widely used tools to access information in electronic networks. On the other, it is clear that search engines provide different kinds of results. These results, in turn, are concentrated among certain sites, and they display both an online structure of their own as well as corresponding in certain respects with various characteristics of the offline world.

Two initial hypotheses have been investigated by means of a combination of quantitative and qualitative methods. First, we have shown that the results for different search engines, in terms of size and recall capacity, are considerably different. This finding was further supported by examining the highest-ranked sites and comparing them with a sample obtained from consultation with experts. This provided validation of the heterogeneity of search engines and indicates that they provide different functionalities in terms of content. Therefore, it is risky to generalize about search engines in general as they differ in their results.

Secondly, there is an embedded social structure inherent in the distribution of resources on the Web. Similar patterns of the centrality of key resources, their connectivity and the way they are clustered, emerge on the Web spaces of the six topics (Climate Change, Internet and Society, Poverty, Trade Reform, Terrorism, and AIDS/HIV). These social networks can be seen as electronic social networks - and making them transparent to users of search engines will be important for the uses of this technology in the future.

Third, to come back to our overall research question, the emerging patterns of Web networks do not uniformly conform to the expectations of a 'winner-take-all' model. Instead, there is some evidence to suggest a 'fragmented' – fractal – structure of information networks. This may be the consequence of the Web tending indeed to reduce the significance of off line hierarchies in accessing information – tending to democratize access around the world. Alternatively, this could indicate a more fine-grained winner-take-all effect whereby centres of expertise progressively refine their specializations and thus gain a 'winner-take-all' status within a narrower area of expertise. While our evidence is not conclusive on either interpretation, it suggests that the winner-take-all model should be viewed as more problematic than heretofore considered.

Finally, there is qualitative information in this study that points to the usefulness of embedding quantitative features of search engines and the results they yield within a more "user-oriented" and topic-specific understanding of these results, which

include peculiar characteristics of "geography and locality," policy considerations influencing the distribution of resources, as well as the more and less academic nature of certain kinds of information on the Web. These characteristics help to explain significant differences in the structure of online networks.

It seems from our qualitative consultations with researchers that there are considerable differences in *how* people make use of search engines, even if it seems that they overwhelmingly confine themselves to the use of the most popular search engine of the day, Google. These preferences in the nature of their searches, what types of search terms are used and how different kinds of searches are combined with the use of other online and offline sources are bound to influence the results of search engines as well as the design of online information systems over the longer term.

The future uses of the functionalities of search engine technologies are still open, but decisions in the early stages of new technologies often have lasting effects. In future research, it will be important to determine more precisely not only which search engines are used by particular groups of researchers, but also whether they confine themselves to the results that they obtain from this source.

Second, it will be of value to investigate what types of search are most common; i.e., whether the use of particular types of search terms by researchers (names of persons, institutions, or topics) reinforces or gets around the fact that searches are heavily biased towards a few results which may follow a power-law distribution and exhibiting cliquishness.

Thirdly, our research has been cross-sectional in nature, taking into account a static snapshot of Web structure and a particular set of searches at a particular point in time. Results provided here should be replicated and extended in the near future using longitudinal data, over a reasonable period of time, which would provide more robust conclusions. This is perhaps the most important refinement of the current research that could be advanced through further research, enabling us to study the marginal bias of the Web toward more or less centralized sources of expertise.

Finally, quantitative results about the structure of the Web and search results need to be integrated much more closely with qualitative research and quantitative research about people's search strategies so that how people search and what they find can be linked. This will eventually allow us to bridge our understanding of the world of online knowledge with what we know about the world of offline knowledge.

## Notes

1 Power laws are observed in many fields, including physics, biology, geography, sociology, or economics. A Power Law relationship between two quantities (e.g. number of Websites and their "popularity") is characterised by a small number of entities in a certain population (i.e. Web sites) representing most of the distribution of a certain variable (i.e. popularity) whereas a large proportion of the population have a reduced weight in the overall distribution. Power law assume the mathematical form of $y = a\,x^{k}$,

where a (the constant of proportionality) and k (the exponent of the power law) are constants.

2 Connectivity and centrality are two characteristics examined when studying the "structure" of networks and social networks. A number of indicators provide ways to "measure" the degree to which nodes are interconnected to each other within a network (connectivity indicators: such as density of the overall network, average distance of the network, etc.). Centrality indicators include the total number of links originating from a certain node (outdegree centrality or activity centrality), the "importance" or "prestige" of a certain node (links pointing to a certain node, indegree), or even the importance of certain nodes in the overall network (flow betweenness) , measuring how central these nodes are to the whole system.

3 The 31 search engines used for the metasearch strategy were: http://search.about.com/, http://www.alexa.com/, http://altavista.com/, http://search.aol.com/, http://askjeeves.com/, http://search.dmoz.org/, http://search.dogpile.com/, http://www.euroseek.com/, http://msxml.excite.com/, http://www.alltheweb.com/, http://findwhat.com/, http://www.google.com/, http://hotbot.lycos.com/, http://search.jayde.com/, http://www.looksmart.com/, http://search.lycos.com/, http://www.mamma.com/, http://www.moonmist.info/, http://search.msn.com/, http://search.netscape.com/, http://srch.overture.com/, http://www.rolist.com/, http://www.scrubtheweb.com/, http://www.search.com/, http://www.searchgate.co.uk/, http://www.searchhippo.com/, http://s.teoma.com/, http://search.thunderstone.com/, http://dpxml.webcrawler.com/, http://www.wisenut.com/, http://search.yahoo.com/.

4 What represents a "Good URL" should in future research assessed by means of interviews and survey data, combined with some kind of expert judgment. At this stage, the criteria for distinguishing "good URLS" from other initial sources, is based upon the existence of explicit references to any of the keywords either on the URL designation itself, or on the title, or the Keyword sections of the Web site.

## References


Albert, R. & Barabasi, A.-L. (2000). Topology of evolving networks: Local events and universality, *Physical Review Letters*, **85**, 5234–5237.

Aquino, S. & Mitchell, T. (2001). Search engines ready to learn. *Technology Review Online*, April 24.

Beaulieu, A. (2005). Sociable hyperlinks: An ethnographic approach to connectivity. In Christine Hine (ed.), *Virtual methods: Issues in social research on the Internet*. (pp.183–97). Oxford: Berg, .

Barabási, A. L. and Albert, R. (1999). Emergence of scaling in random networks. *Science*, **296**: 509–512.

Benkler, Y. (2006). *The wealth of networks*. New Haven: Yale University Press.

Borgmann, C. & Furner, J. (2002). Scholarly communication and bibliometrics. In B Cronin (ed), *Annual review of information science and technology*, (pp.3-72). Medford, N.J.: Information Today.

Brin, S. & Page, L. (1998). The anatomy of a large-scale hypertextual Web search engine. In *Proceedings of the Seventh International World Wide Web Conference* (pp.107 – 117). Amsterdam: Elsevier.

Broder, A. (2002). Taxonomy of Web search. *ACM SIGIR Forum archive*. **36**(2), Fall 2002.

Chau, M., Chen, H., Qin, J., Zhou, Y., Qin, Y., Sung, W. & McDonald, D. (2002). Comparison of two approaches to building a vertical search tool: A case study in the nanotechnology domain. *Proceedings of the 2nd ACM/IEEE-CS Joint Conference on Digital Libraries*:, 135 – 144

Dutton, W. H. (2005), The Internet and social transformation: Reconfiguring access. In Dutton, W. H., Kahin, B., O'Callaghan R., & Wyckoff, A. W. (eds.), *Transforming enterprise* (pp. 375–97), Cambridge, MA: MIT Press.

Dutton, W. H., & Helsper, E. (2007), *The Internet in Britain 2007*. Oxford: Oxford Internet Institute, University of Oxford. Available at: http://www.oii.ox.ac.uk/research/oxis/OxIS2007_Report.pdf

Dutton, W. H., di Gennaro, C., & Millwood-Hargrave, A. (2005), *The Internet in Britain*. Oxford Internet Institute, University of Oxford. Available at: http://www.oii.ox.ac.uk/research/oxis/oxis2005_report.pdf

Eastman, C. M. & Jansen, B. J. (2003). Coverage, relevance, and ranking: The impact of query operators on Web search engine results. *ACM Transactions on Information Systems*, **21**(4), October 2003.

Fallows, D. (2005). Search engine users. Pew Internet and American Life Project. Washington, DC. (available online www.pewInternet.org).

Flake, G., Lawrence, S. & Giles G. (2000). Efficient identification of Web communities. *Proceedings of the Sixth International Conference on Knowledge Discovery and Data Mining (ACM SIGKDD-2000)*, Boston, MA: 150 – 160.

Flake, G., Lawrence, S., Giles, C. & Coetzee, F. (2002). Self-organisation and identification of Web communities, *Computer Magazine*, March 2002, 66 – 71.

Fortunato, S., Flammini, A., Menzcer, F., & Vespignani, A. (2006). Topical interests and the mitigation of search engine bias. *Proceedings of the National Academy of Sciences*, **103**: 12684–12689.

Frank, R. H., & Cook, P. J. (1995), *The Winner-Take-All Society*. New York: The Free Press.

Fry, J. (2004). Scholarly research and information practices: a domain analytic approach, *Information Processing and Management*, **42**: 299–316.

Fry, J., Virkar, S., & Schroeder, R. (2008). Search engines and expertise about global issues: Well-defined landscape or undomesticated wilderness, In A. Spink & M. Zimmer (eds.) *Websearch: Interdisciplinary Perspectives*. London: Springer.

Fry, J. & Talja, S. (2004). The cultural shaping of scholarly communication: Explaining e-Journal use within and across academic fields, in *Proceedings of the American Society for Information Science and Technology Annual Meeting on Managing and Enhancing Information: Cultures and Conflicts*. (pp.20–30), Providence, Rhode Island, 13-18 November 2004.

Hargittai, E. (2004). Do you "Google"? Understanding search engine use beyond the hype, *First Monday*, issue 9(3). (online journal).

Hellsten, I., Leydesdorff, L. & Wouters, P. (2006). Multiple presents: How search engines re-write the past. *New Media and Society*, **8**(6), 901–924.

Hess, D. (1997). *Science studies: An advanced introduction*. New York: New York University Press.

Huang, L. (2002), *A survey on Web information retrieval technologies*. State University of New York at Stony Brook. Accessed via HTTP June 6, available at http://citeseer.nj.nec.com/336617.html

Kleinberg, J. & Lawrence, S. (2001). The Structure of the Web, *Science*, **294**: 1849 – 1850

Kleinberg, J. M. (1999). Authoritative sources in a hyperlinked environment. *Journal of the ACM (JACM)* **46**(5): 604–632.
Kleinberg, J. (1998). Authoritative sources in a hyperlinked environment. *Proceedings of the ACM-SIAM* Symposium on Discrete Algorithms.
Kumar, R., Raghavan, P., Rajalopagan, S., Sivakumar D., Tomkins, A. S. Upfal, E. (2000). The Web as a graph. *Proceedings of the 19th Symposium on Principles of Database Systems*, p. 1
Lawrence, S. & Giles C. (1999). Accessibility of information on the Web. *Nature*, **400** (6740), 107–109.
Matzat, U. (2004). Academic communication and Internet discussion groups: transfer of information or creation of social contacts?, *Social Networks*, **26**, 221–255.
Merton, R. K. (1968). The Matthew Effect in science. *Science*, **159**(3810): 56–63.
Park, H. W. & Thelwall, M. (2005). The network approach to Web hyperlink research and its utility for science communication. In Christine Hine (ed.), Virtual Methods: Issues in Social Research on the Internet. (pp.171–81). Oxford: Berg.
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*, **99**(8), 5207–5211.
Persson, O. & Beckmann, M. (1995). Locating the network of interacting authors in scientific specialties. *Scientometrics*, **33**(3), 351–366
Ravasz, E. & Barabasi, A. L. (2003) "Hierarchical organization in complex networks." *Physical Review* E67 026112.
Spink, A. &Jensen, B. J. (2004). *Web search: Public searching of the Web*. London: Springer.
Walsh, J. P.; Kucker, S.; Maloney, N. G. & Gabbay, S. (2000). Connecting minds: Computer-mediated communication and scientific work. *Journal of the American Society for Information Science*, **51**(14), 1295–1305.

## About the Authors

Alexandre Caldas is Director of CEGER (The Management Centre for the Electronic Government Network) in Portugal, Research Associate of the Oxford Internet Institute in the UK and Visiting Assistant Professor of University Atlântica. His research is focused on Information and Communication Technologies, Webmetrics, Internet indicators, e-Science and e-Government, and social networks. alexandre.caldas@ceger.gov.pt
**Address:** Management Center for the Electronic Government Network, Rua Almeida Brandão 7, 1200-602 Lisbon, Portugal

Ralph Schroeder is James Martin research fellow at the Oxford Internet Institute at Oxford University. His current research interests include the social implications of e-Science and social interaction in virtual environments. ralph.schroeder@oii.ox.ac.uk
**Address:** Oxford Internet Institute, 1 St Giles, Oxford OX1 3JS UK

Gustavo S. Mesch is an Associate Professor in the Department of Sociology & Anthropology at the University of Haifa (Israel). His research interests include inequalities of access in the Information society, the effects of computer mediated communication on social networks and social capital. gustavo@soc.haifa.ac.il

**Address:** Department of Sociology & Anthropology, University of Haifa, Har Hacarmel 31905, Israel

**William H. Dutton** is Director of the Oxford Internet Institute, Professor of Internet Studies at the University of Oxford, and a Professorial Fellow at Balliol College. His research is focused on the social dynamics of the Internet, from everyday life to research environments. william.dutton@oii.ox.ac.uk
**Address:** Oxford Internet Institute, 1 St Giles, Oxford OX1 3JS UK