

What Can We Learn About Distributed Problem Solving Networks Via Web Analysis?

Robert Ackland

Australian Demographic and Social Research Institute, The Australian National University, robert.ackland@anu.edu.au

1 Introduction

The mode of Internet-enabled organization referred to here as “distributed problem-solving networks” (DPSNs) is also known as bazaar governance (Demil and Lecocq, 2006; Hope, 2008), commons-based peer production (Benkler, 2006), crowd sourcing, collective wisdom and wkinomics (Tapscott and Williams, 2007). There is currently debate in the academic literature as to how this mode of organizing relates to other governance structures: hierarchy (where collaboration is centrally coordinated through managers), market (where coordination is via price signals) and network (where coordination is via relationships between actors). Demil and Lecocq (2006) and Hope (2008), among others, regard the bazaar as a discrete, fourth, form of governance structure for coordinating transactions.

This purpose of the present paper is not to contribute to the theory of bazaar governance and whether it should be considered as something different to network coordination (with which it shares some conceptual similarities, as evidenced, for example, by the fact that Benkler’s book is titled “The Wealth of Networks”). Rather, we focus on to what extent network analysis, and in particular, analysis of hyperlink data from the public Web, can shed light on the performance of various DPSNs featured as case studies in the Oxford Internet Institute-McKinsey Technology Initiative (OII-MTI) project titled “Performance of Distributed Problem-Solving Networks”.

Our finding is that for the majority of the DPSNs we consider, network analysis of data from the public Web does not provide direct insights into performance. In fact, in our opinion, data from the public Web can currently only be used to provide direct performance metrics for one type of entity or organization featured in the OII-MTI project – distributed news aggregators (Richter, Escher, and Bray, 2008) – and a preliminary web analysis of news aggregators is provided in this paper.

With regards to the other DPSNs featured in the OII-MTI project, we argue that public Web data can only provide indirect, but still useful, insights into the functioning and performance of organizations or entities.

2 Analyzing the performance of DPSNs using data from the public Web

Our focus here is on the role of empirical or quantitative analysis of public Web data, by which, we mean data that are freely available either via using web crawlers to extract website text and hyperlinks or via APIs into search engines such as Google and Yahoo. Broadly, there are three approaches for conducting empirical analysis of Web data, each reflecting a particular disciplinary base (we acknowledge that this is a fairly crude categorization, and we acknowledge that there is in fact active cross-over between the various approaches).

First, “webometrics” (also known as webmetrics and cybermetrics) is an approach for analyzing hyperlink data and website usage patterns that largely draws on bibliometrics and infometrics (see, for example, Almind and Ingwersen, 1997; Björneborn and Ingwersen, 2004; Thelwall, Vaughan, and Björneborn, 2005).

Second, applied physicists have focused on the identification of empirical properties in large-scale collections of Web pages and the development of statistical-mechanical models that can be used to explain the emergence of such properties. For example, Newman (2002) has studied the existence of “assortative mixing” (or correlation between the attributes of adjacent network nodes) in large-scale networks and found that the Web exhibits *disassortative* mixing (or degree anti-correlation between nodes), reflecting the fact that high-degree web pages are often directories that by definition tend to connect to low-degree individual web pages. Barabási and Albert (1999) explain the existence of “power laws” in the distribution of links in large-scale networks such as the WWW, where the “rich get richer” and a small number of sites receive the lion’s share of links pointing toward them, via the concept of preferential attachment: newer entrants are inclined to link to already well-connected actors, thereby increasing these incumbents’ advantage.

The final broad approach to empirical analysis of the Web has emerged from the social sciences and features statistical analysis of hyperlink networks (using approaches from social network analysis) in order to understand the role of the Web in enabling various forms of social, economic or political behaviour. For example, research has been conducted into the formation of hyperlink networks by political parties (Ackland and Gibson, 2004) and environmental social movement organizations (Ackland, O’Neil, Bimber, Gibson, and Ward, 2006; O’Neil and Ackland, 2006). Shumate and Dewitt (2008) statistically analyze the hyperlink network of 248 HIV/AIDS NGOs, identifying “structural signatures” that reflect that the NGOs are engaging in collective action behaviour by contributing to the formation of an “information public good”. González-Bailó (2007) employs social network analysis of hyperlink networks formed by civil society actors to understand the role of Web in promoting plurality and empowering the poor. In the remainder of this paper (an in particular, in Section 3) our focus will be primarily on this third approach to analyzing Web data.

To understand the role of data from the public Web in analyzing the performance of DPSNs, we first classified the OII-MTI case studies into three groups.¹

1 Note that there are three other case studies in the OII-MTI project that we do not study here: open source software development (Dalle, den Besten, Masmoudi, and David, 2008), distributed film production (Cassarino and Geuna, 2008), and the Large Hadron Collider ATLAS Experiment (Tuertscher, 2008). The reason is that in contrast to the case studies we do study, the major activity being undertaken by the actors in these DPSNs does not occur via the use of Web applications.

2.1 Web 2.0 business applications

The first group is what we term “Web 2.0 business applications”. Sermo is a startup company with a business model centers on hosting a social networking site for US doctors, and then charging pharmaceutical and healthcare companies for the right of participating in the community in order to crowd-source possibly valuable medical information (see Bray, Croxson, Dutton, and Konsynski, 2008b). Seriosity is a startup company that aims to sell office productivity applications that employ principles drawn from online gaming environments (see Bray, Croxson, Dutton, and Konsynski, 2008a).² Finally, information markets are markets (generally, Web-based) where individuals can buy and sell assets whose payoff is tied to an outcome in the future, such as the winner of the US Democratic Primary race (see Bray, Croxson, and Dutton, 2008). The business model for information markets either involves selling software for setting up intra-organizational markets or hosting an online information market (and selling the crowd-sourced information to third parties).

With regards to Sermo, public Web data are unlikely to shed much light on the performance of this DPSN. While website traffic is an indication of performance (if US doctors are continuing to make use of the service in large numbers, then this indicates that they find it useful as a means of discussing problems and finding answers), such data would be closely-held by the company. Website traffic might similarly provide some insight into the performance of online information markets, but again, this would be difficult to obtain. In addition, website traffic data would only be really that useful (as a metric of performance of a DPSN) if traffic statistics could be compared across other sites providing a similar service.

If there were enough examples of particular type of online DPSN (thus warranting an analysis of website traffic data for example, to see if there is a correlation between characteristics of DPSNs and their performance, as measured by website traffic), then there are three approaches for gaining comparative website data. Hitwise provide measures of traffic that involve analysis of weblog files of participating Internet Service Providers, while Alexa produce user-centric measures of traffic (they collect data on web browsing habits of individuals who have installed an Alexa plugin in their web browser). A third approach that has been suggested for estimating the relative popularity of web applications involves periodically polling local Domain Name servers (LDNSs) that service Domain Name Server requests (Wills, Mikhailov, and Shang, 2003). When a user accesses a website (such as Sermo) their LDNS converts the Sermo URL into its IP address and the URL-IP mapping will be stored in its cache for a set period of time (since LDNS requests for cached URLs can be served much quicker than if there needs to be a follow-on request to another LDNS). Regular probing of the DNS cache will provide indirect information on how often particular URLs are accessed by users of the LDNS. While this approach is clever and has potential, it is obviously complicated and we are not aware of anyone currently providing this as a service to researchers.

However, even website traffic data may not provide an entirely accurate picture of the relative performance of DPSNs such as Sermo and online information markets. This is because people do not always use the best website (or buy the best product, for that matter) – they are often influenced by marketing and “buzz”, or what their friends use. The economics of superstars (Rosen, 1981) has been used to explain why in certain fields (e.g.

² Currently, the main Seriosity product is Microsoft Outlook email plugin, which is strictly not an example of Web 2.0. However, Seriosity is clearly embracing (or intending to embrace) Web 2.0 technologies in its approach, with the co-founder of the company (Byron Reeves) arguing that “...embracing emerging communications and Web 2.0 technologies gives leaders more ways to communicate, which in turn allows them to be more effective communicators.” (http://www.seriosity.com/downloads/GIO_PDF_web.pdf)

the arts and sport) there is a concentration of output among a few talented individuals, and an associated marked skewness in the distribution of income, with very large rewards at the top. For Adler (1985), an interesting puzzle about stardom is that stars are often not more talented than many artists who are less successful, and he presents a model where large differences in earnings can exist with *no differences* in talent.

The relevance of this for our analysis of the performance Web 2.0 business applications is that one of the pre-conditions for the emergence of superstars (according to the models of Rosen and Adler) is that there exists a “joint consumption technology” where the cost of production does not rise in proportion to size of a seller’s market. The Web is an example of a joint consumption technology – the fixed costs associated with setting up a website such as Sermo would greatly outweigh the cost of serving up individual web pages, and hence average costs per client would fall as the client base rises. The implication is that superstar websites can (and do) emerge, and the economics of superstars thus provides an economic rationale for why power laws exist on the Web. New startup Web business thus have a great incentive to maximize their visibility on the Web by ensuring that their new website is being mentioned by influential tech bloggers, for example, since this will help to push up their ranking on search engines such as Google (the search engine ranking of a given site is heavily influenced by the number of relevant web pages that link to it) which will translate into website traffic.

2.2 Wikipedia

The second group of case studies includes just one example, Wikipedia, which is a free online encyclopedia constructed by the efforts of thousands of individuals (den Besten, Loubser, and Dalle, 2008). Public Web data could possibly be useful for assessing the performance of this DPSN; it would be possible to use hyperlink network analysis to ascertain the prominence or visibility of particular pages on the Wikipedia site (for example, pages on prominent and well-defined topics) and make a comparison with the visibility of pages on competing sites such as Encyclopedia Britannica. To the extent that people link to pages containing definitions that they find useful and accurate, then this hyperlink analysis could provide some insights into performance. Website traffic to the pages would again provide useful information (but again, these data would be hard to obtain).

2.3 News aggregators

As mentioned above, we argue that the performance of online news aggregator sites (Richter, Escher, and Bray, 2008) *can* be directly analysed using data from the public Web; in the next section this is demonstrated via a preliminary analysis of a sample of online media sites.

3 The performance of online media websites

A preliminary analysis of the performance of online media websites is now presented. The approach is based on methods for analyzing Web data that have been developed as part of

the Virtual Observatory for the Study of Online Networks (VOSON) project ³ and are detailed further in Ackland and Gibson (2004), Ackland, O’Neil, Bimber, Gibson, and Ward (2006) and O’Neil and Ackland (2006).

3.1 Dataset construction

A sample of 40 websites was selected for this preliminary analysis (these are referred to below as “seed sites”). The websites were chosen to be representative of four types of online media actors: news aggregators, bloggers, social bookmarking sites and traditional media sites. For the blogger sites, “tech” bloggers were further distinguished from “other” bloggers. Note that while the social bookmarking sites do not necessarily focus on news, they contribute to searchability on the web, and hence it was decided to include them in the preliminary analysis. The 40 seed sites are listed in Table 1.

The aim of the data collection process is to find the hyperlink connections between the 40 seed sites. Since many of the seed sites are huge it was not technically feasible to crawl them (and some of the sites such as digg.com prevent crawler access via the robots protocol). It was therefore decided to use the Yahoo API to find inbound hyperlinks to each of the seed sites. The Yahoo API provides an estimate of the total number of hyperlinks pointing to a given domain (site_in_e in Table 1). The social bookmarking site <http://del.icio.us/> has the highest (estimated) number of inbound links (nearly 10 million) - this is almost 2.5 times the estimated inbound link count for the next highest (the tech blogger <http://www.engadget.com/>).

The Yahoo API will only provide the first 1000 inbound links to a given domain. These inbound hyperlinks were collected for each of the 40 seed sites (site_in_a in Table 1 shows the actual number of hyperlinks returned from the API and it is apparent that this was less than the maximum of 1000 for five of the sites). A problem with the data collection approach is that we do not know how Yahoo determines the order in which hyperlinks are returned. It is expected that the returned hyperlinks are ordered in terms of relevance or authority (using an internal Yahoo ranking scheme) and this has implications for our data collection approach (and the accuracy of the results presented below). Since the number of hyperlinks returned is truncated at 1000 this may mean that there is a bias toward finding connections between the sites that Yahoo deem to be more prominent (in that they are returned earlier) and a bias towards not finding connections between the less prominent sites. A sampling procedure might be needed to address this problem.

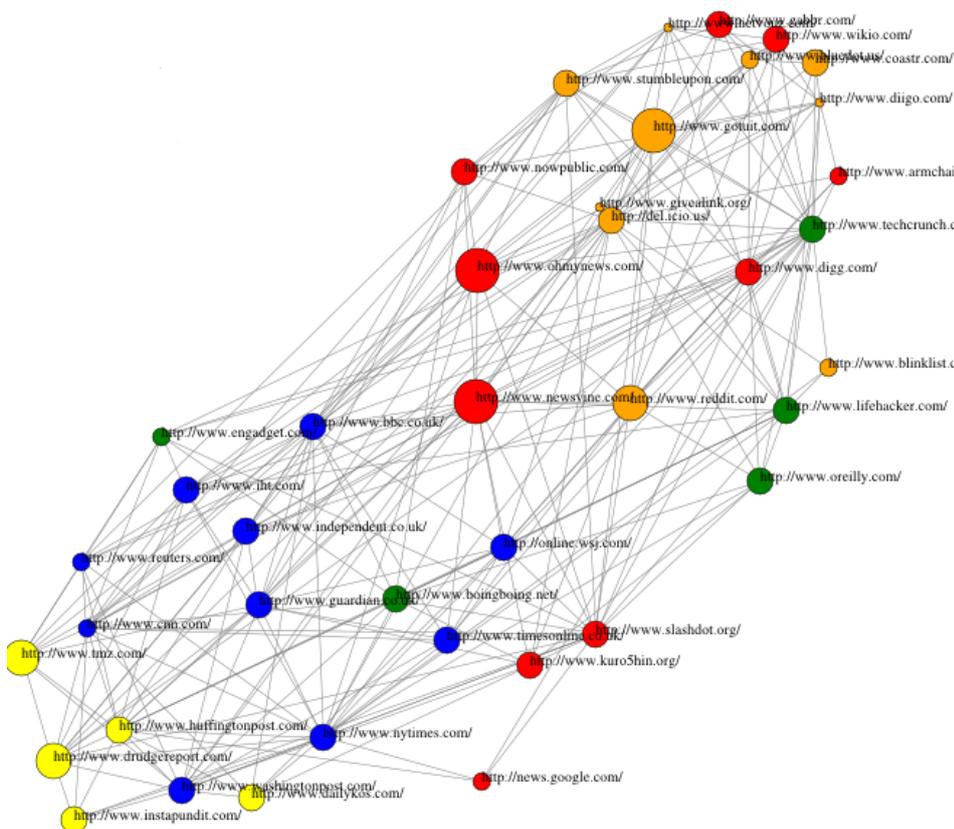
The above data collection resulted in a database containing almost 40,000 web pages. Various automated data cleaning and processing steps were then undertaken (detail on these is omitted here for brevity), with the major steps being the grouping of all pages from a given domain and the further grouping of pages from multiple domains that are managed by or related to a single organization. For example, the network node for slashdot.org includes pages from nearly 20 domains, including: <http://www.slashdot.org/>, <http://linux.slashdot.org/>, <http://yro.slashdot.org/> and <http://politics.slashdot.org/>. The pages_n column in Table 1 shows the number of pages included in the database for each of the seed sites.

While the database contains the URLs of many web pages that were linked to the seed sites, in this analysis we focus only on the links between the seed sites. A subset database containing only the 40 seed sites was then created, and the network map for this subset is shown in Figure 1 (node size is proportional to indegree). The map is drawn using the LinLogLayout force-directed graphing (FDG) algorithm of Noack (2005) which is designed for

3 <http://voson.anu.edu.au>

mapping small-world graphs such as those found on the WWW.⁴ The force-directed approach for identifying clusters in graphs is well established. In the context of our web graph, web sites are given initial random positions and modeled as electrostatic charges (repulsion forces that act to push nodes apart from one another). Hyperlinks between web sites are modeled as springs (attraction forces that act to pull together those sites that are connected to one another via hyperlinks). The algorithm shifts the position of nodes in an attempt to minimize the energy of the system (in general, the energy of the system will be smaller if two connected nodes are positioned near one another compared with if they are on separate sides of the map). The process of moving nodes around the map to minimize the energy of the system can reveal web clusters or communities - collections of sites that have more links to other members in the collection than to nodes outside the collection.

Figure 1: Force-directed graph for 40 seed sites [blog-other (yellow), blog-tech (green), news aggregator (red), social bookmarking (orange), traditional media (blue)]



There is reasonably clearly defined clustering in the map: the traditional media and “other” (i.e. non-tech) bloggers are one side, while the social bookmarking and 3 of the 5 tech bloggers are on the other side. Two of the news aggregator sites (newsvine.com and ohmynews.com) appear to occupy central positions between both “camps”.

4 See <http://www.informatik.tu-cottbus.de/an/GD/> for more details.

3.2 Node-level measures: Characterizing online network positions of media sites

Figure 1 presents a directed graph – the media sites are nodes and hyperlinks between the sites are edges.⁵ If we take the view that each media website is an actor in a socially-generated network (i.e. the web as a network of actors and not just a network of documents), then social network analysis (SNA) may provide insights into the roles of various actors in the online media sector and also metrics for performance.

Table 1 shows indegree and outdegree for the 40 seed sites (remember that this is just hyperlinks between the seed sites). The average indegree over all 40 sites is 6.8 (i.e. the average seed site has nearly 7 hyperlinks from other seeds pointing to it) – social bookmarking sites have below average indegree (5.5) while news aggregator sites have above average indegree (7.5). The average outdegree over all 40 sites is 6.8 (this must equal the indegree given its a closed system) – news aggregator sites have an average of only 4.1 while tech bloggers have an average of 10.2 and traditional media sites have an average of 9.2. The high average outdegree for traditional media sites was surprising and perhaps indicates the extent to which traditional media has “caught up” and embraced the web in terms of hyperlinking to other media sites. Note that with a sample of only 40 sites the standard errors for these averages (and those that follow) will be very high, so the conclusions need to be treated with caution.

Table 2 shows the counts of hyperlinks between different types of media sites. Of the 92 hyperlinks that traditional media sites made (to other seed sites), over a third went to blog sites. News aggregator sites were the biggest recipient of hyperlinks from other media sites (75 of the 271 hyperlinks). Tech bloggers were much more inclined to link to social bookmarking and news aggregator sites, compared with other bloggers.

Table 1 also reports summary statistics for measures calculated using the Hyperlinked-Induced Topic Search (HITS) algorithm of Kleinberg (1999), which is based on the premise of the existence of hyperlinked communities that contain two distinct, but inter-related, types of pages – authorities (highly referenced pages) on the topic, and pages that point to the authority pages. The latter are referred to as hubs since they serve as central points from which authority is conferred on other authority pages in the community. Thus, there is a mutually reinforcing relationship between authorities and hubs: a good hub points to many good authorities, and a good authority is pointed to by many good hubs. To calculate the HITS measures, each pagegroup p in the component is associated with an authority weight $x(p)$ and a hub weight $y(p)$, which are initialised to 1. In a single iteration of HITS $x(p)$ is replaced by the sum of the y 's of all pagegroups pointing to p , and $y(p)$ is replaced by the sum of the x 's that page p points to. After each iteration, the x 's and y 's are normalized and convergence is generally achieved after less than 10 iterations.

From Table 1, the average hub score (calculated over all 40 sites) was 0.117 and the average authority score was 0.151. News aggregator sites appear to have a lower-than-average hub score (0.078), while traditional media have an above average hub score (0.158). Again, this is surprising (and it reflects the similar finding for outbound hyperlinks) since the expectation was that news aggregator sites would be pointing users to quality authority sites which would be the traditional media. However, it should be noted again that the small sample is problematic here, and three of the news aggregator sites (newsvine, digg and slashdot) have above average hub scores. As a group, the tech bloggers have the highest hub score but this is all because of one site <http://www.techcrunch.com/>

⁵ Note that the arrow heads are not shown in this figure.

The final measures in Table 1 are network betweenness and closeness. Betweenness centrality gives an indication of the extent to which an individual node plays a “brokering” or “bridging” role in a network and is calculated for a given node by summing up the proportion of all minimum paths within the network that “pass through” the node. Closeness centrality is an indicator of the extent to which a given node has short paths to all other nodes in the graph and it is thus a reasonable measure of the extent to which the node is in the “middle” of a given network.

To the extent that the purpose of new online media actors such as bloggers (and in particular, news aggregators) is to help people to find useful and relevant information on the web, betweenness centrality, in particular, may provide a useful measure of the relative performance of particular actors. Nodes with high level of betweenness tend to promote “searchability” of the network and also help to bring together various clusters or communities of nodes that otherwise might not have much connection with one another. In terms of betweenness centrality, the standout performers are <http://www.newsvine.com> and <http://www.techcrunch.com>, with scores of 153.4 and 111.7 respectively, compared with an overall average of 31.8.

3.3 Network-level measures: Characterizing the networks formed by online media sites

Network-level measures have not been calculated in this preliminary report, however we propose that they may be useful for constructing relative performance measures of particular *classes* of online media actors (e.g. bloggers, compared with news aggregators).

Centralization is a network-level property that is calculated for a given node-level property, and it broadly measures the distribution of importance, power or prominence among actors in a given network. In effect, centralization measures the extent to which the network “revolves around” a single node or small number of nodes and the classic example of a highly centralized network is the star network where the node in the Centre of the star has complete centrality while the other nodes have minimal centrality (this network has the maximum centralization score of 1). In contrast, a circle network is highly decentralized since all nodes share the same centrality and such a network will have a centralization score of 0.

Network structure influences how rapidly information spreads throughout the network. In a dense network, information will tend to flow more quickly because path lengths will on average be lower. Also centralized networks are more efficient in transferring information since the central nodes will be able to connect information seekers to the appropriate sources of the information. Centralized networks therefore exhibit “searchability” – when information is distributed around a network, people will know that there are certain sites that they can visit that will connect them to these sources of information. However, highly-centralized networks can also be more vulnerable in the sense that if a central node is removed, then the performance of the network will suffer.

3.4 Online media actors – future work

This represents a very preliminary insight into how insights into the performance of online media actors may be gained via the application of social network analysis to hyperlink data. Future work would need to involve analysis of larger samples of online media actors. Further, the fact that the Yahoo API only returns a maximum of 1000 hyperlinks may introduce bias into the analysis; the nature and direction of this bias needs to be analysed, and approaches for correction developed. Finally, it would be useful to conduct analysis of the text on the web

pages that have been collected – via text analysis, it may be possible to identify the types of stories that are contributing to various network properties e.g. searchability.

References

- Ackland, R., and R. Gibson (2004): "Mapping Political Party Networks on the WWW," Paper presented at the Australian Electronic Governance Conference, 14-15 April 2004, University of Melbourne.
http://acsr.anu.edu.au/staff/ackland/papers/political_networks.pdf
- Ackland, R., M. O'Neil, B. Bimber, R. Gibson, and S. Ward (2006): "New Methods for Studying Online Environmental-Activist Networks," paper presented to 26th International Sunbelt Social Network Conference, 24-30 April, Vancouver.
http://voson.anu.edu.au/papers/environmental_activists_methods.pdf Accessed 26th February, 2008.
- Adler, M. (1985): "Stardom and Talent," *American Economic Review*, 75(1), 208–212.
- Almind, T. C., and P. Ingwersen (1997): "Informetric analyses on the World Wide Web: Methodological approaches to 'webometrics'," *Journal of Documentation*, 55(4), 404–426.
- Barabási, A.-L., and R. Albert (1999): "Emergence of Scaling in Random Networks," *Science*, 286, 509–512.
- Benkler, Y. (2006): *The Wealth of Networks: How Social Production Transforms Markets and Freedom*. Yale University Press, New Haven.
- Björneborn, L., and P. Ingwersen (2004): "Toward a basic framework for webometrics," *Journal of the American Society for Information Science and Technology*, 55(14), 1216–1227.
- Bray, D., K. Croxson, and W. Dutton (2008): "Information Markets: Feasibility and Performance," Working paper for OII-MTI "Performance of Distributed Problem-Solving Networks" project. http://www.oii.ox.ac.uk/research/dpsn/Informationmarkets_brief.pdf Accessed: 26th February, 2008.
- Bray, D., K. Croxson, W. Dutton, and B. Konsynski (2008a): "Seriosity: Addressing the Challenges of Limited Attention Spans," Working paper for OII-MTI "Performance of Distributed Problem-Solving Networks" project.
http://www.oii.ox.ac.uk/research/dpsn/Seriosity_brief.pdf Accessed: 26th February, 2008.
- Bray, D., K. Croxson, W. Dutton, and B. Konsynski (2008b): "Sermo: A Community-Based Knowledge Ecosystem," Working paper for OII-MTI "Performance of Distributed Problem-Solving Networks" project.
http://www.oii.ox.ac.uk/research/dpsn/Sermo_brief.pdf Accessed: 26th February, 2008.
- Cassarino, I., and A. Geuna (2008): "Distributed film production: Artistic experimentation or feasible alternative?," Working paper for OII-MTI "Performance of Distributed Problem-

- Solving Networks” project. http://www.oii.ox.ac.uk/research/dpsn/ASOA_brief.pdf
Accessed: 26th February, 2008.
- Dalle, J.-M., M. den Besten, H. Masmoudi, and P. David (2008): “Bug-Patching for Mozilla’s Firefox,” Working paper for OII-MTI “Performance of Distributed Problem-Solving Networks” project. http://www.oii.ox.ac.uk/research/dpsn/Firefox_brief.pdf Accessed: 26th February, 2008.
- Demil, B., and X. Lecocq (2006): “Neither Market nor Hierarchy nor Network: The Emergence of Bazaar Governance,” *Organization Studies*, 27(10), 1447–1466.
- den Besten, M., M. Loubser, and J.-M. Dalle (2008): “Distributed Problem Solving in Wikipedia,” Working paper for OII-MTI “Performance of Distributed Problem-Solving Networks” project. http://www.oii.ox.ac.uk/research/dpsn/Wikipedia_brief.pdf Accessed: 26th February, 2008.
- González-Bailó, S. (2007): “Mapping Civil Society on the Web: Networks, Alliances, and Informational Landscapes,” DPhil Thesis, University of Oxford.
- Hope, J. (2008): *Biobazaar: The Open Source Revolution and Biotechnology*. Harvard University Press, Cambridge, MA.
- Kleinberg, J. (1999): “Authoritative Sources in a Hyperlinked Environment,” *Journal of the ACM*, 46(5), 604–632.
- Newman, M. E. J. (2002): “Assortative Mixing in Networks,” *Phys. Rev. Lett.*, 89, 208701.
- Noack, A. (2005): “Energy-Based Clustering of Graphs with Nonuniform Degree,” in *Proceedings of the 13th International Symposium on Graph Drawing*. GD 2005, Limerick, Sep. 12-14.
- O’Neil, M., and R. Ackland (2006): “The Structural Role of Nanotechnology-Opposition in Online Environmental-Activist Networks,” paper presented to 26th International Sunbelt Social Network Conference, 24-30 April, Vancouver.
http://voson.anu.edu.au/papers/environmental_activists_structural_role.pdf Accessed 26th February, 2008.
- Richter, W., T. Escher, and D. Bray (2008): “The Performance of Distributed News Aggregators,” Working paper for OII-MTI “Performance of Distributed Problem-Solving Networks” project. http://www.oii.ox.ac.uk/research/dpsn/Newsaggregators_brief.pdf
Accessed: 26th February, 2008.
- Rosen, S. (1981): “The Economics of Superstars,” *American Economic Review*, 71(5), 845–858.
- Shumate, M., and L. Dewitt (2008): “The North/South Divide in NGO Hyperlink Networks,” *Journal of Computer-Mediated Communication*, 13, 405–428.
- Tapscott, D., and A. Williams (2007): *Wikinomics: How Mass Collaboration Changes Everything*. Atlantic Books, London.
- Thelwall, M., L. Vaughan, and L. Björneborn (2005): “Webometrics,” *Annual Review of Information Science and Technology*, 39, 81–135.

Tuertscher, P. (2008): "The ATLAS Collaboration – A Distributed Problem-Solving Network in Big Science," Working paper for OII-MTI "Performance of Distributed Problem-Solving Networks" project. http://www.oii.ox.ac.uk/resrarch/dpsn/Atlas_brief.pdf Accessed: 26th February, 2008.

Wasserman, S., and K. Faust (1994): *Social Network Analysis: Methods and Applications*. Cambridge University Press, Cambridge.

Wills, C., M. Mikhailov, and H. Shang (2003): "Inferring Relative Popularity of Internet Applications by Actively Querying DNS Caches," IMC'03, October 27-29, 2003, Miami Beach, Florida, USA.

Table 1: Online network measures for media sites

url	type	indeg	outdeg	pages_n	site_inlinks_e	site_inlinks_a	hits_hub	hits_auth	closenes s	betweenness
http://www.digg.com/	news aggregator	6	11	18	2720000	1000	0.192	0.131	0.549	26.234
http://www.newsvine.com/	news aggregator	11	10	15	905000	1000	0.201	0.231	0.459	153.377
http://www.slashdot.org/	news aggregator	6	16	39	487000	1000	0.302	0.130	0.609	32.498
http://www.kuro5hin.org/	news aggregator	8	1	3	133000	1000	0.015	0.164	0.146	5.901
http://www.armchairgm.com/	news aggregator	5	0	3	1030000	1000	0.000	0.124	0.143	0.000
http://www.gabbr.com/	news aggregator	7	0	3	2430	919	0.000	0.147	0.143	0.000
http://www.nowpublic.com/	news aggregator	8	1	4	35100	1000	0.029	0.166	0.250	1.043
http://www.wikio.com/	news aggregator	8	0	5	17300	1000	0.000	0.167	0.143	0.000
http://english.ohmynews.com/	news aggregator	12	2	7	170000	1000	0.041	0.281	0.328	77.287
http://news.google.com/	news aggregator	4	0	1	2010000	1000	0.000	0.125	0.143	0.000
Average		7.5	4.1	9.8	750983	991.9	0.078	0.167	0.291	29.634
http://www.washingtonpost.com/	traditional media	7	11	31	1310000	1000	0.194	0.146	0.549	36.102
http://news.bbc.co.uk/	traditional media	7	15	32	1210000	1000	0.267	0.159	0.582	55.772
http://www.reuters.com/	traditional media	5	7	28	2450000	1000	0.113	0.130	0.542	13.849
http://www.cnn.com/	traditional media	5	16	30	3550000	1000	0.280	0.089	0.619	59.978

http://www.nytimes.com/	traditional media	8	18	41	3460000	1000	0.290	0.196	0.591	74.421
http://www.iht.com/	traditional media	8	6	11	354000	1000	0.104	0.206	0.488	45.484
http://www.guardian.co.uk/	traditional media	8	11	24	2190000	1000	0.201	0.158	0.542	81.306
http://www.timesonline.co.uk/	traditional media	6	4	9	330000	1000	0.073	0.160	0.481	10.772
http://online.wsj.com/	traditional media	8	4	7	53200	1000	0.059	0.175	0.476	39.486
http://www.independent.co.uk/	traditional media	8	0	4	862000	1000	0.000	0.165	0.143	0.000
Average		7	9.2	21.7	1576920	1000	0.158	0.158	0.501	41.717

Note: indeg - number of inbound hyperlinks from other seeds; outdeg - number of outbound hyperlinks to other seeds; pages_n - number of web pages in database for this site; site_inlinks_e - total number of inbound hyperlinks (as estimated by Yahoo); site_inlinks_a - number of hyperlinks actually collected (note: max. is 1000); hits_hub - HITS hub score (see text); hits_auth - HITS authority score (see text); closeness (see text); betweenness (see text).

Table 1: Online network measures for media sites (cont.)

url	type	indeg	outdeg	pages_n	site_inlinks_e	site_inlinks_a	hits_hub	hits_auth	closenes s	betweenness
http://www.engadget.com/	tech blogger	5	5	16	3900000	1000	0.071	0.100	0.506	16.972
http://www.techcrunch.com/	tech blogger	8	22	41	1260000	1000	0.357	0.129	0.696	111.703
http://www.boingboing.net/	tech blogger	6	7	15	1210000	1000	0.104	0.138	0.464	14.430
http://www.lifehacker.com/	tech blogger	7	10	17	491000	1000	0.146	0.136	0.557	78.952
http://radar.oreilly.com/	tech blogger	6	7	17	151000	1000	0.101	0.160	0.513	15.359
Average - tech blogger		6.4	10.2	21.2	1402400	1000	0.156	0.133	0.548	47.483
http://www.huffingtonpost.com/	other blogger	7	15	11	1730000	1000	0.251	0.119	0.565	56.045
http://www.dailykos.com/	other blogger	7	0	2	1760000	1000	0.000	0.141	0.143	0.000
http://www.drudgereport.com/	other blogger	10	9	3	1450000	1000	0.146	0.234	0.506	63.465
http://www.tmz.com/	other blogger	9	4	9	1830000	1000	0.062	0.169	0.453	35.721
http://www.instupundit.com/	other blogger	6	1	3	1490000	1000	0.024	0.152	0.339	0.000
Average - other blogger		7.8	5.8	5.6	1652000	1000	0.097	0.163	0.401	31.046
Average		7.1	8	13.4	1527200	1000	0.126	0.148	0.474	39.265
http://www.bluedot.us/	social bookmarking	4	5	12	1980	666	0.105	0.114	0.269	10.829
http://del.icio.us/	social bookmarking	6	20	31	9920000	1000	0.341	0.118	0.672	89.773
http://www.diigo.com/	social bookmarking	3	9	18	791000	1000	0.169	0.088	0.527	4.118

http://www.givealink.org/	social bookmarking	1	0	1	752	202	0.000	0.035	0.143	0.000
http://www.netvouz.com/	social bookmarking	3	8	14	675000	1000	0.148	0.089	0.513	4.826
http://www.reddit.com/	social bookmarking	9	7	13	72100	1000	0.133	0.166	0.345	34.980
http://www.blinklist.com/	social bookmarking	4	0	1	2510000	1000	0.000	0.118	0.143	0.000
http://www.coastr.com/	social bookmarking	6	0	3	838	346	0.000	0.137	0.143	0.000
http://www.gotuit.com/	social bookmarking	13	0	1	2000	953	0.000	0.280	0.143	0.000
http://www.stumbleupon.com/	social bookmarking	6	9	16	2320000	1000	0.162	0.161	0.494	22.319
Average		5.5	5.8	11	1629367	816.7	0.106	0.131	0.339	16.684
Average - All sites		6.8	6.8	14.0	1371118	952	0.117	0.151	0.402	31.825

Table 2: Counts of hyperlinks between different types of media sites

.	blog-other	blog-tech	news aggregator	social bookmarking	traditional media	All
blog-other	10	2	2	0	15	29
blog-tech	2	11	18	16	4	51
news aggregator	4	3	13	11	10	41
social bookmarking	2	4	27	16	9	58
traditional media	21	12	15	12	32	92
All	39	32	75	55	70	271

Annex: Some social network analysis measures

There are several social network analysis (SNA) measures that are used in this paper – these are briefly explained here (for more details on social network analysis, the standard text is Wasserman and Faust, 1994).

A *network* is a set of nodes (or vertices) and a set of ties (or edges) indicating connections between the nodes. The relational ties in a network may be “directed” (e.g. person X knows person Y , but person Y may not know person X) or “non-directed” (e.g. if person X has a familial relationship with person Y , the converse must also be true). Relational ties may also be “dichotomous” (e.g. person X either has a familial relationship with person Y or doesn't) or “valued” (e.g. in a network representing trading patterns between countries, the value of a network tie between countries X and Y may be the value of exports as a percentage of GDP).

Node-level measures

The *indegree* of a node is the sum of ties that the node receives from other nodes in the network, while *outdegree* is the sum of ties that the node makes to other nodes in the network. In the context of hyperlink networks, indegree is the number of hyperlinks directed towards the site (or web page) while outdegree is the number of hyperlinks on the website that point to other sites.

The *betweenness centrality* measures the extent to which a node is positioned on the shortest path (or “geodesic”) between other pairs of nodes in the network. The betweenness centrality $C_B(v)$ for vertex v is:
$$C_B(v) = \sum_{s \neq v \neq t \in V(G)} \frac{\sigma_{st}(v)}{\sigma_{st}}$$
 where $\sigma_{st}(v)$ is the number of shortest paths from node S to t that pass through v , and σ_{st} is the total number of shortest paths from node S to t . Betweenness is sometimes normalized to sum to 1 over all nodes. For a 3-node network, if actor 1 linked to actor 2 and actor 2 linked to actor 3, then the betweenness centrality of actor 2 would be 1 (since there is 1 minimum path from actor 1 to actor 3, and it passes through actor 2), while the betweenness for actors 1 and 3 would be zero.

Closeness centrality is an indicator of the extent to which a given node has short paths to all other nodes in the graph and it is thus a reasonable measure of the extent to which the node is in the “middle” of a given network. The closeness centrality $C_C(v)$ for vertex v is:
$$C_C(v) = \frac{|V(G)-1|}{\sum_{i:i \neq v} d(v,i)}$$
 where $d(v,i)$ is the geodesic distance from v to i (where defined). Closeness is sometimes normalized to sum to 1 over all nodes. For our 3-node network, if actors 1 and 2 both linked to one another, and actors 2 and 3 both linked to one another, then the

closeness score for actor 2 would be 1, while for actors 1 and 3 the centrality score would be 2/3.

Network-level measures

Inclusiveness is the number of connected actors as a proportion of the network size.

Density is the number of ties between nodes as a proportion of the maximum number of ties.

Dyadic reciprocity is the number of mutual dyads (pairs of nodes with reciprocated links) as a proportion of the total number of dyads (pairs of linked nodes where the links are either unilateral or reciprocated), while *edge reciprocity* is the proportion of edges which are reciprocated.

Centralization is a network-level property that is calculated for a given node-level property, and it broadly measures the distribution of importance, power or prominence among actors in a given network. Centralization is calculated by first calculating a particular centrality measure and then finding the sum of the absolute deviations from the graph-wide maximum (generally, the centralization score is also normalized by the theoretical maximum centralization score). For a given graph G and given centrality measure $C(v)$, where v is a node in graph G , the unnormalized centralization score is:

$$C^*(G) = \sum_{i \in V(G)} \left| \max_{v \in V(G)} (C(v)) - C(i) \right|$$

To illustrate the calculation of indegree centralization, consider a three-node star network where actors 2 and 3 both link only to actor 1, and actor 1 doesn't link to anyone. The unnormalized indegree centralization score is $(2-2) + (2-0) + (2-0) = 4$, which (given we are working with a star network) is equal to the theoretical maximum centralization score and hence the normalized indegree centralization score is 1. At the other extreme is a circle network where actor 1 links to actor 2, actor 2 links to actor 3 and actor 3 links to actor 1. The unnormalized indegree centralization score is $(1-1) + (1-1) + (1-1) = 0$, and the normalized indegree centralization score is also 0.

The centralization scores for outdegree and indegree are calculated analogously to above, however betweenness and closeness centrality require additional explanation. For a 3-node network, if actor 1 linked to actor 2 and actor 2 linked to actor 3, then the unnormalized betweenness centralization score for this network is $(1-0) + (1-1) + (1-0) = 2$, while the normalized betweenness centralization score is 0.5. Note that this is calculated as the unnormalized betweenness centralization score (2) divided by the theoretical maximum centralization score for a 3-node network (4) which is calculated for a directed network with the formula $(N-1)^2 * (N-2)$, where N is the number of nodes in the network.

For our 3-node network, if actors 1 and 2 both linked to one another, and actors 2 and 3 both linked to one another, then the unnormalized closeness centralization score for this network is $(1-2/3) + (1-1) + (1-2/3) = 2/3$, while the normalized closeness centralization score is 0.5. Note that this is calculated as the

unnormalized closeness centralization score (2/3) divided by the theoretical maximum centralization score for a 3-node network (4/3) which is calculated for a directed network with the formula $(N - 1) * (1 - 1/N)$, where N is the number of nodes in the network.