Discussion Paper: Network and Concept Maps for the Blogosphere

Dr Axel Bruns	*QUT Creative Industries*	*a.bruns@qut.edu.au*
Lars Kirchhoff	*Universität St. Gallen*	*lars.kirchhoff@unisg.ch*
Thomas Nicolai	*Universität St. Gallen*	*thomas.nicolai@unisg.ch*
Dr Jason Wilson	*QUT Creative Industries*	*j5.wilson@qut.edu.au*
Barry Saunders	*QUT Creative Industries*	*b.saunders@qut.edu.au*
Tim Highfield	*QUT Creative Industries*	*t.highfield@qut.edu.au*

0. Status of This Document

Discussion paper version 1.0 (1 May 2008). First draft by Axel Bruns, edited and discussed by the group; for internal and external discussion.

Further intentions:

- send to other stakeholders (Citizen Journalism Linkage project: Terry Flew, Stuart Cunningham, Graham Young – also Martin Stewart-Weeks?; John Hartley; Robert Ackland of VOSON; also Richard Rodgers of IssueCrawler, people from Leximancer and QUT High Performance Computing?)
- post on *Gatewatching.org* for public discussion
- use as basis for the ISEA 2008 and AoIR 2008 papers by the group

1. Background

The blogosphere allows for the networked, decentralised, distributed discussion and deliberation on a wide range of topics. Based on their authors' interests, only a subset of all blogs will participate in any one topical debate. Even within such debates, there will be an uneven distribution of participation based on a variety of sociocultural factors:

- the time available for any individual blogger to participate,
- the blogger's level of interest in the topic,
- the blogger's awareness of other blogs discussing the topic (which they may link and respond to),
- the blogger's status amongst their peers (which may determine how aware others are of the blog, and thus whether they will read, comment on, link to, or respond to the blogger's posts),
- the quality of the blogger's writing and contributions,
- the blogger's specific interests in the topic (which may lead them to focus on particular aspects of the wider topic),
- and additional factors including the blogger's political ideology, gender, age, location, sociodemographic status (to the extent that these are evident from the blog), as well as the language they write in.

In combination, these factors mean that networked debate on specific topics in the blogosphere is characterised by clustering (Barabási, Albert & Jeong, 1999; Newman, Watts & Strogatz, 2002; Watts, 1999). For any one topic, there are likely to be one or multiple clusters of highly active and closely interlinked blogs, surrounded by a looser network of blogs which are less active contributors to the debate and are less densely linked to it. Individual clusters in the topical debate may be able to be distinguished according to certain factors: for example, their topical specialisation (focussing on specific sub-topics of the wider debate) or their shared identity (e.g. a common national, ethnic, or ideological background).

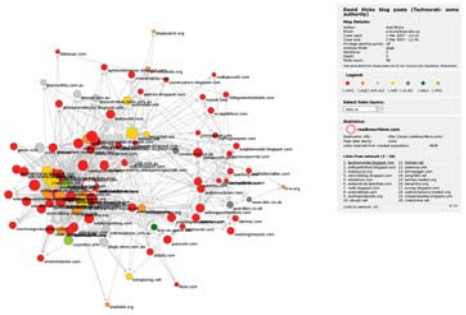

Fig. 1: Network of Bloggers in the Australian Political Blogosphere (from Bruns, 2007)

Such blog-based debate is difficult to conceptualise under the general terms of the Habermasian public sphere model (which as formulated depends on the existence of a dominant mass media to ensure that all citizens are able to be addressed by it); at a smaller level, however, it may be possible to understand networked discussion on specific topics in the blogosphere to constitute what may be described as a public spherule (Bruns 2008). Rather than seeing networked political debate in terms of the operations of a public sphere, we can think about a group of topical discussion clusters of sufficient size and interconnection providing a substitute for their participants. It may be that when layered on top of one another, the public spherules on various topics of public interest can stand in as a replacement for the conventional public sphere (whose existence is undermined by the decline of the mass media *as* mass media; see Castells 2007). This *networked* public sphere would necessarily be more decentralised than the conventional, Habermasian model of the public sphere.

2. Research Framework

While qualitative evidence for the networked patterns of discussion, debate, and deliberation in the blogosphere is readily available, it is more difficult to establish a solid quantitative picture of blog-based topical discussion networks and their cluster patterns. Large numbers of blogs (and individual blog posts, links, and comments) are likely to be involved in a quantitative study of blog-based discussion patterns. Hence, automated data collection and analysis is necessary. Any tools used for this purpose need to be able to distinguish between the different units of analysis:

- **blog content** must be divided into a number of functional categories:
    - **blog:** the Website on which blog posts are published – a multi-page collection of all current and archived blog posts and comments, *including* other ancillary information (blogger information, blogroll, navigation bars, ads, etc.). Identified by a generic URL, e.g. http://www.myblog.com/ or http://blogger.com/myblog/
    - **blogroll:** a common blog design element which provides a more or less static list of other blogs and Websites which the blog owner finds useful (not related directly to the topic of individual blog posts).
    - **blog post:** an individual blog entry – usually published on a single Web page, but *not including* any ancillary information which may also exist on that page (such as header, footer, and sidebars). Identified by a specific URL, e.g. http://www.myblog.com/posts/734
    - **blog comments:** reader comments submitted in response to a blog post – usually published on the same page as the post itself, below the post (also *not including* other comment on that page, such as the post itself, other comments, or other ancillary comments). Identified by a specific URL, e.g. http://www.myblog.com/posts/734#comment-23

- **links** between Websites must also be distinguished according to their function:
    - **generic links:** links on blogs which exist outside of blogrolls, blog posts, or blog comments – including navigational links, links to search engines, Web standards organisations, or links in on-site ads.
    - **blogroll links:** links to other blogs and related Websites which are of lasting interest to the blog author. Usually linking to generic URLs, e.g. http://www.otherblog.com/ or http://blogger.com/otherblog/
    - **topical links:** links included in blog posts and blog comments, usually pointing to specific blog entries or other Web pages which are of direct relevance to the content of the blog post. Often (but not exclusively) pointing to specific URLs, e.g. http://www.otherblog.com/posts/234
    - **commenter-provided links:** links from the names or nicknames of commenters to their own blogs or Websites. Usually linking to generic URLs, e.g. http://www.otherblog.com/ or http://blogger.com/otherblog/

Distinctions between these different categories build on the following assumptions:

1. Content:

The core underlying assumption is that *the vast majority of bloggers write about topics which interest them* (rather than claiming an interest they don't have). This should not be understood to claim that bloggers cover *all* the topics they are interested in, however – the topics covered on any one blog constitute merely that subset of all interests which a blogger has deemed it acceptable to reveal publicly to a general readership. On this basis, we assume that:

a. The complete collection of all blog posts for a given blog provides a reliable indication of the interests of the individual blogger (as expressed publicly); the development of these interests may be further traced by tracking changes in topical coverage over time.
b. A comparison of bloggers' interests (in total, for specific periods of time, and/or in relation to broad topical domains) across multiple blogs indicates the distribution of topical interest across the blogosphere (at least for the subset of the entire blogosphere included in the analysis).
c. A comparison between the blogger's postings on specific topics, and the collection of reader comments to these postings, indicates the level of agreement or disagreement between blogger and commenters (at least for blogs with substantial commenting activity).

2. Links:

The core underlying assumption is that *links to other Websites indicate a recognition of the linked content as 'interesting'* (for a variety of possible reasons, and potentially indicating approval or disapproval). By extension, *this also confers a certain amount of reputation and attention on the creator of the linked content* (again,this accrued reputation can be either positive or negative).

We also assume that linking patterns predict traffic and influence. The more incoming links any piece of content has, the more likely visitors are to see it, and this increases its potential to influence readers. Further, the outgoing links of sites which themselves receive many incoming links are more powerful in directing traffic and conferring influence than the outgoing links of little-known sites. *Google*'s PageRank and *Technorati*'s authority ranking operate on similar assumptions.

On this basis, we assume that *patterns of interlinkage indicate the existence of a network of attention. These patterns are indicators of visibility and influence. In these patterns, the balance of incoming and outgoing links for any one site or page warrants special attention*. Specifically,

a. Patterns of interlinkage between contemporaneous blogrolls indicate the existence of a long-term network of recognition between peers. Sites with many incoming *and* outgoing links may be understood as *hubs* for communication in this network; sites with many incoming, but limited outgoing links may be understood as central *sources* for information; sites with many outgoing but few incoming links may be understood as (not necessarily central) *distributors* of attention to other members of the network. The gradual evolution of such networks can be traced over time.
b. Patterns of interlinkage between contemporaneous blog posts (and other post-level content) indicate the existence of a network of debate on specific topics. Such networks of debate can be seen to persist over greater or lesser periods of time. Posts with many incoming links may be understood as making an important (possibly controversial) *original* contribution to the debate; posts with many incoming and outgoing links may be understood as making an important *discursive* contribution to the debate; posts with many outgoing links may be understood as *introductions to* or *summaries of* ongoing debate.
c. Aggregated from the level of the blog post to that of the blog, these patterns of interlinkage also indicate the role of the overall blogs in topical debate networks. Blogs with many incoming *and* outgoing links may be understood

as central *hubs* for communication on this topic; blogs with many incoming, but limited outgoing links may be understood as central *sources* for information on the topic; blogs with many outgoing but few incoming links may be understood as (not necessarily central) *distributors* of attention to other members of the network. A comparison of these short-term debate networks over time and across topics indicates the fluctuation of centrality; sites whose centrality remains high over time can be seen as having significant authority overall, while sites whose centrality is high only for specific topics can be seen as having significant authority only for those topics.

d. Patterns of interlinkage between blog posts and comments indicate that posts or comments have an ongoing relevance to particular networked debates. If a comment is linked to in a further post (either on the blog on which the comment was posted, or elsewhere), it indicates that the comment has itself provoked further discussion and commentary, and that the conversation constitutes a dialogue between blogosphere authors and commenters. If blog posts are referred to in comments threads, especially if these are on other blogs, it indicates that the initial post has relevance and influence in an ongoing, networked debate.

e. Patterns of linkage between current and archived posts on the same blog indicate the blog author's continuing interest in and coverage of relevant topics.

3. Research Methodology

In order to conduct a quantitative analysis of blog-based discussion networks at a content and link level, a number of tools must be used. Each introduces a number of necessary limitations to the breadth and depth of study possible. The three key elements of the research process are data gathering and processing, content analysis, and link network analysis (however, this does not imply that content analysis necessarily precedes network analysis, or vice versa). Subsequently, it is also possible to extract and identify common patterns and interrelations between content und network analyses. Additional work beyond these initial stages could extend into social network analysis, to identify social networks within the Blogosphere.

Data Gathering and Processing

Blog content of interest to this project is openly available on the Web (content on blogs behind intranet firewalls and password protected blogs cannot be regarded as being part of public discussion as we define it here). Further, most blogs offer RSS feeds which alert subscribers to new posts. RSS feeds in themselves are an insufficient data source, however: some contain only excerpts from whole posts, and many do not contain links, images, or other functional elements of the blog posts. None contain comments (though separate RSS feeds for comments to a specific blog post may also be available).

For a full and reliable analysis, it is therefore necessary to scrape entire blog pages with all textual and functional elements. This, however, also creates problems as it will include the site's navigational elements, blogrolls, comments, ads, and other ancillary material in the data gathered. A direct blog post-level analysis of the data will therefore produce skewed results.

This means that scraped blog pages must be further processed in order to separate the salient content (the blog posts itself) from ancillary material; in the process, other salient elements (blogrolls, comments) can also be gathered and stored in separate categories. Such processing is non-trivial and time-consuming. Further, page layout and formatting is inconsistent across blogs, and the scraped data processor must be trained for each category or sometimes for individual blogs (for example, although Wordpress and Blogger software are commonly used, there are many different versions and templates possible).

If comments are to be considered, there is an important choice to be made as to *when* scraping will occur. When an RSS feed announces the publication of a new blog post, few if any comments will have been made. Many more comments are likely to have been offered an

hour, a day, or a week after the publication of the post. Later comments may be less relevant to the immediate discussion For these reasons, a fixed period between the publication of a blog post and the scraping of post and comments will therefore need to be set.

For practical reasons, and unless direct access to the up-to-date page archives of a commercial search engine is available, the number of blogs scraped will also need to be limited; it is not feasible to scrape the entire blogosphere, or even a large part of it. Instead, our methodology must content itself with focussing on a specific and manageable part of the blogosphere – for example, Australian political blogs. Even here, a comprehensive coverage may be impossible. It is possible that Australian political blogs exist which are so little-known and unconnected that they are invisible in standard sources like *Google* and *Technorati*. And generally non-political or non-Australian blogs may contain a very occasional post about Australian politics, but fall outside the scope of the study. Nevertheless, coverage of a large part of Australia's political blogosphere is possible, with the core rather than the far periphery of the network is the focal point of analysis.

Even here, though, the list of blogs (and related sites) to be scraped should be viewed as open and growing, and to be established over multiple iterations of the scraping process:

0. First, a seed list of well-known blogs in the field is compiled by the researchers.
1. Content on these seed blogs is scraped.
2. Outgoing links in the scrape data are identified, and relevant blogs and related sites are added to the scrape list.
3. Steps 1 and 2 are repeated at least until the rate by which new blogs are added to the list slows substantially, and/or until new sites identified during scraping are demonstrably outside of the core field of interest studied here.
4. RSS feeds for each of the blogs are identified, and new posts are scraped as they become available (and/or with a built-in delay to allow for the accumulation of comments attached to posts).
5. The scraping process continues.

Content Analysis

Content analysis builds on the data gathered in the scraping process, operating on the level of blog posts (or blog posts plus blog comments). It uses automated large-scale quantitative content analysis tools such as Leximancer to identify terms, themes, and concepts in the data (or in subsets of the entire corpus of data), and their interrelationships. Such automated content analysis should be further followed up by reading selective posts and comments in a more qualitative examination of specific issues, concepts or conversations.

Potential approaches to content analysis include:

a. *Determination of overall key terms, themes, and concepts across the entire corpus.*
   This provides an overview of the core themes of the specific section of the blogosphere under analysis here. Beyond short-term fluctuations, it can be seen as a picture of the stable and recurring interests of the overall population of blogs studied; this can then also be usefully compared to similar studies conducted for different sections of the global blogosphere (e.g. comparing Australian and British political bloggers, etc.) (references to such studies).
b. *Change of themes over time.*
   Using the same approach as a. above, this determines key terms, themes, and concepts for a specific sub-set of the corpus, as defined by the time of posting (e.g. week by week, or for specific periods of interest, such as pre/post-election). This identifies the level of correlation between blog-based discussions and wider public debate in the mainstream media and elsewhere; additionally, it may also indicate how new themes and keywords spread from one or a handful of original posts across the wider blogosphere (and what role central hubs in the blog network play in this process).
c. *Identification of key themes for individual bloggers or groups of blogs.*

Rather than focussing on a specific time period, this identifies key themes for specific bloggers or groups of blogs in the corpus. Groups may be chosen for example according to network clustering tendencies as identified through network analysis (see below), or in relation to other identifying factors. Such analysis may also be further broken down into specific time periods (bloggers for whom the content analysis identifies an interest in specific topics may be expected to be particularly active at times when such topics are central to public debate, for example, and this analysis would be able to identify such patterns). By conducting such analyses for different groups of blogs within the overall sample, it would be possible to identify their differing topical interests; where content analysis shows frequent use of specific keywords, underlying ideological patterns may also be able to be identified and distinguished.

d. *Comparison of commenters' and bloggers' content*.
A comparison of the key terms, themes and concepts present in the corpus of blog posts and the corpus of blog comments (again possibly sliced according to timeframe, bloggers and blogger groups, and/or other factors) would indicate the specific level of conceptual consistency between bloggers and commenters. At the level of individual blogs as well as for groups of broadly like-minded blogs, this would indicate the extent to which site users make comments mainly on blogs with whose views they agree (that is, it would help quantify the extent to which blogs act as echo chambers). At the same time, it should be noted that this analysis does not make any predictions about the likelihood of agreement or disagreement by non-commenting readers or by other bloggers linking to a blog post from their own site.

e. *Comparisons of treatment of key issues between particular blogs and blog communities, or between clusters of blogs.*
"Deep" comparisons of the treatment of specific issues in the blogosphere can be carried out by reading specific bloggers' posts and the comments that follow. This can augment quantitative analysis by showing the terms in which conversations between posters and commenters are carried out in particular blogs on particular issues, and how intangibles like tone vary across specific blogs or clusters of blogs. Such analysis can also reveal how consensus is developed on blogs, how norms are enforced (e.g. treatment of trolls and newcomers), and whether particular issues excite more vociferous debate.

Some interesting work is being done in this field by IBM in analysing communication streams in real time (Spangler, Kreulen et al. 2006).

Network Analysis

Network analysis focusses on the network of interlinkages between blogs at blogroll, blog post, and blog comment levels. It uses automated large-scale network analysis tools such as VOSON to trace the networks of interlinkage and identify clusters of closely interlinked nodes in the network, distinguishing also between inlinks and outlinks.

Potential approaches to link network analysis include:

a. *Identification of static networks of blogs using blogroll links.*
As a form of *acknowledgement* of blogs and other related Websites, blogrolls indicate not so much the day-to-day interconnection of blogging activity, but longer-term connections between bloggers. Considered specifically, blogroll links can thus be used to identify relationships between blogs and bloggers *as bloggers themselves see them*. (This is necessarily limited by the extent to which bloggers use blogrolls.) Sites with a high number of incoming links are particularly interesting here, as these can be understood as the most respected blogs in the overall population. Additionally, sites with high inlinks *and* outlinks can be seen as important *hubs* in the network, helping to make connections between otherwise relatively insular clusters in the network. Further analysis of changes in the blogroll network over time also shows how blogs wax and wane in their peers' esteem over time.

b. Identification of discursive networks on specific issues using blog post links.

By selecting a specific period of posts, network analysis on a blog post level can identify the discursive networks existing at that particular time. This can help identify the most influential *original* posts at the time (most inlinks) and the most *discursively* productive posts (most inlinks *and* outlinks). Further manual and automated content analysis of these and/or all posts during this time period can further distinguish different topical clusters which may overlap with the network clusters identified in the network analysis (see content analysis section, c., above, and combination analyses section below).

c. *Identification of discursive networks on specific topics above the level of blog posts*.
By attributing all individual posts on a specific blog to that blog, the analysis described in b. above can be repeated on a more general level to identify the discursive networks between blogs as such for a specific timeframe (and by extension, for the topics on which blog-based discussion was focussed during that timeframe). This produces not an image of the highly changeable day-to-day discursive interlinkage between individual blog posts, but a somewhat more stable network of communicative interlinkage between blogs. Further content analysis can be performed for individual clusters in the resulting network to identify any conceptual differences between them.

d. *Identification of general and specific discussion leaders.*
By further extending c. above to study fluctuations in blog networks over time, it is possible to identify which blogs in the overall population are *continuously* central to blog networks, and which are central *only at specific times* (further content analysis of blog posts on these specific blogs would likely show a difference in their topical interests). Aggregated over time, such patterns can also be compared with the blogroll network examined through a. above in order to establish whether the bloggers' self-proclaimed patterns of static connection reflect the patterns of discursive interconnection found here.

Combination Analyses

There are many opportunities for correlations between conceptual and network analyses (and for further triangulation using additional sources, including closely reading posts and threads, comparison with information about key themes in the mainstream media during specific timeframes, and correlation with site rank indicators such as *Google*'s PageRank or *Technorati*'s authority index). Indeed, neither content nor network analyses in isolation provide a detailed picture of the blogosphere; there is a need to augment one with the other and with other data.

Possible combination analyses include:

a. *Relating network fluctuations to changing topical focus*.
Results from a network analysis of discursive networks in the blogosphere can be correlated with the issues of the day as identified by content analysis for the same time period. Network analysis can also be compared with various analyses of the mainstream media. Analyses can also be done over different time periods, such that fluctuations in the shape of the network as different events may serve to highlight various specialisations and interests across the blogosphere.

b. *Correlating network and concept clusters*.
Clusters of closely interlinked blogs and posts, as identified through network analysis (as described in network analysis section, b. and c., above), may correlate with clusters of conceptual themes uncovered by content analysis (see content analysis section, c., above). Content analysis of the core blogs in a given network cluster, or network analysis of the posts in a given content cluster (for both cases, at first for a limited timeframe to avoid overgeneralisation) should show such correlations if they exist. That is to say that different network clusters are expected to have have a different conceptual focus, and different concept clusters should have a different population of blogs participating in them. Where similar groups of blogs are repeatedly found to exhibit clustering tendencies, identifying their specific conceptual preoccupations would help avoid the use of more simplistic labels (e.g. 'labour rights

blogs' and 'climate change-sceptic blogs' rather than simply 'left/right-of-centre blogs').

c. *Identifying distinguishing features of core blogs.*
Content analysis that focusses less on topical key terms and more on stylistic features of blog texts may be able to contribute an answer to what has made core blogs in the network successful. Comparison between a group of core blogs and a group of peripheral blogs may be able to identify stylistic differences (writing level and style, form of address to readers, as well as other factors such as number of inlinks and outlinks, frequency of posts, etc.).

d. *Correlation with external measures of site rank.*
Concept and network analyses can be further correlated with external site data including *Google* PageRank and Technorati authority scores. This will provide a useful reality check for these scores. This will indicate whether the core sites in the networks (as determined by network analysis) are also higher in these external rankings. It may help identify the role which highly ranked sites play in disseminating concepts which originate with lower-ranked blogs in the network (this would be an extension of the approach described in the content analysis section, b., above). As global indices, PageRank and authority scores may turn out to be overly vague predictors of authority for the limited subset of the blogosphere under consideration.

Further opportunities for combined analyses may be identified during the course of our research. Generally, all analysis models outlined above may be deepened through close readings of blogs, in addition to the automated methods on which this methodology builds.

4. Limitations

A number of limitations apply for this research programme and have been already identified above:

- For practical reasons, analysis is necessarily limited to a subset of the overall blogosphere. It may miss aspects of the data which exceed the limits of the group of blogs studied here. Thus, the quality of findings from analysis is likely to be better nearer the core of the concept and link networks than it is on the periphery.
- For some of the analytical approaches outlined above, additional limitations may need to be introduced (e.g. selecting a specific timeframe or a specific cluster of blogs for in-depth study). Such limitations suffer from similar border issues, and repeated analysis with differently defined limitations may need to be performed to compare outcomes and optimise the methodological approach.
- The identification of concept and network clusters makes certain assumptions about what constitutes a cluster (that is, to what degree the correlation of terms and the interlinkage of sites are indicative of close clustering). Experimentation with cluster definitions and with various measures of proximity may be necessary to compare outcomes and optimise the methodologies discussed.

5. Applications of this Research

The research methodology described above can be applied in a variety of ways, for various purposes. Broadly, it will enable researchers to

- indicate the shape of the networked public sphere overall, and of the individual public spherules which we assume may constitute it;
- show the level of polarisation of or interconnection between the participants in public debate within any such public spherule;
- indicate similarities and differences between various subsets of the overall blogosphere, as defined for example by topic, nationality, or, language;
- track the evolution and dissemination of individual memes (terms, themes, concepts) across the blogosphere, thereby providing a quantitative basis for the application of extant communications theories to communication in the blogosphere;
- show evidence of the collective knowledge distributed across the blog network.

(Further detail to be added here.)

Barabási, A.-L., Albert, R., & Jeong, H. 1999. “Scale-Free Characteristics of Random Networks: The Topology of the World-Wide Web.” In Physica A 281, pp. 69-77.

Bruns, A. 2008. “Life beyond the Public Sphere: Towards a Networked Model for Political Deliberation.” In Information Polity 13 (1-2), pp. 65-79.

———. 2007. “Methodologies for Mapping the Political Blogosphere: An Exploration Using the IssueCrawler Research Tool.” In First Monday 12 (5). http://www.uic.edu/htbin/cgiwrap/bin/ojs/index.php/fm/article/view/1834/1718 (accessed 29 Apr. 2008).

Castells, M. 2007. “Communication, Power and Counter-Power in the Network Society.” In International Journal of Communication 1, pp. 238-266.

Habermas, J. 2006. “Political Communication in Media Society: Does Democracy Still Enjoy an Epistemic Dimension? The Impact of Normative Theory on Empirical Research.” In Communication Theory 16 (4), pp. 411-26.

Leximancer. 2008. http://www.leximancer.com/ (accessed 29 Apr. 2008).

Newman, M. E. J., Watts, D. J., & Strogatz, S. 2002. “Random Graph Models of Social Networks.” In PNAS 99 (1), pp. 2566-2572.

Spangler, W., J. Kreulen, et al. (2006). "Machines in the Conversation: Detecting themes and trends in informal communication streams." IBM Systems Journal **4**(45): 785.

VOSON: Virtual Observatory for the Study of Online Networks. 2008. http://voson.anu.edu.au/ (accessed 29 Apr. 2008).

Watts, D. J. 1999. “Networks, Dynamics, and the Small-World Phenomenon.” In The American Journal of Sociology 105 (2), 493-527.