Locating the Australian Blogosphere: Towards a New Research Methodology

Dr Axel Bruns, Dr Jason Wilson, Barry Saunders, Tim Highfield
Creative Industries Faculty, Queensland University of Technology, Brisbane, Australia
a.bruns, j5.wilson, b.saunders, t.highfield@qut.edu.au

Lars Kirchhoff, Thomas Nicolai
Institute for Media and Communication Management, University of St. Gallen, Switzerland
lars.kirchhoff, thomas.nicolai@unisg.ch

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.

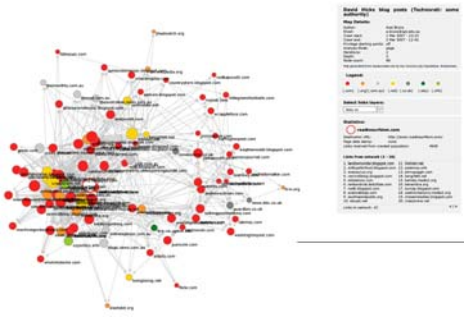

Fig. 1: Network of Bloggers in the Australian Political Blogosphere (from Bruns, 2007)

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).

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; see Habermas 2006); 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.

Our project aims to develop a rigorous and sound methodology for the study of this networked 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: in terms of content, the blog posts themselves, blog comments, blogrolls, and ancillary (static) content; in terms of links, topical links in blog posts, commenter-provided links, blogroll links, and generic links elsewhere on the site.

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).

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.

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 (2008) 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.
b. Change of themes over time.
c. Identification of key themes for individual bloggers or groups of blogs.
d. Comparison of commenters' and bloggers' content.
e. Comparisons of treatment of key issues between particular blogs and blog communities, or between clusters of blogs.

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 (2008) 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.
b. Identification of discursive networks on specific issues using blog post links.
c. Identification of discursive networks on specific topics above the level of blog posts.
d. Identification of general and specific discussion leaders.

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.
b. Correlating network and concept clusters.
c. Identifying distinguishing features of core blogs.
d. Correlation with external measures of site rank.

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.

Our presentation at ISEA2008 will demonstrate this research approach in practice, and showcase early findings from an exploratory study of the Australian political blogosphere.

References

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.

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.