Locating the Australian Blogosphere: Towards a New Methodology

Dr Axel Bruns
Creative Industries Faculty
Queensland University of Technology, Brisbane
a.bruns@qut.edu.au

Dr Jason Wilson
Creative Industries Faculty
Queensland University of Technology, Brisbane
j5.wilson@qut.edu.au

Barry Saunders
Creative Industries Faculty
Queensland University of Technology, Brisbane
b.saunders@qut.edu.au

Tim Highfield
Creative Industries Faculty
Queensland University of Technology, Brisbane
t.highfield@qut.edu.au

Lars Kirchhoff
Institute for Media and
Communication Management
University of St. Gallen, CH
lars.kirchhoff@unisg.ch

Thomas Nicolai
Institute for Media and
Communication Management
University of St. Gallen, CH
thomas.nicolai@unisg.ch

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, with varying intensity, based on a variety of sociocultural factors: a blogger's time, interest, and awareness of current discussion; their status in the blogosphere; the topical focus of their contributions; and their political ideology, gender, age, location, sociodemographic status, 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). 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).

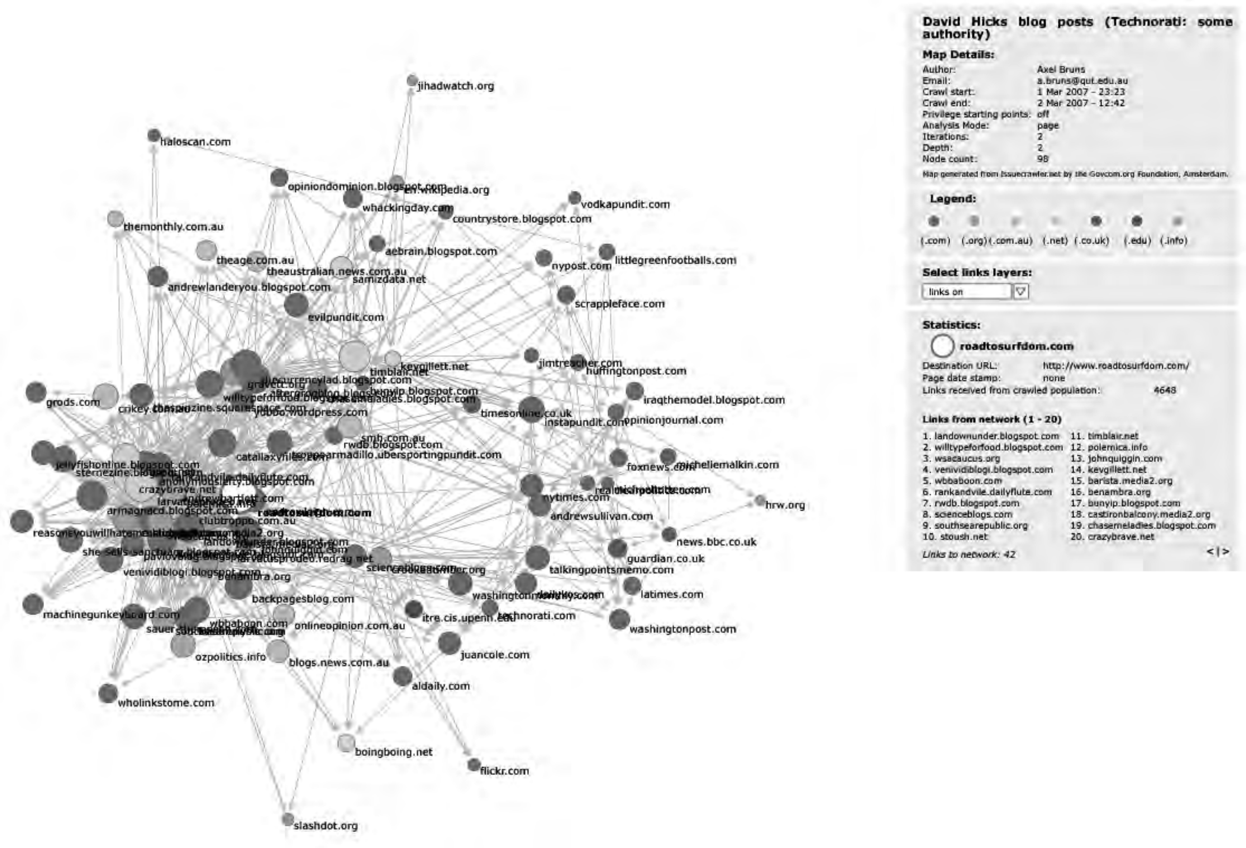
Such blog-based debate is difficult to conceptualise under the general terms of the Habermasian public sphere model (which as formulated depends on the existence of a dominant mass media to ensure that all citizens are able to be addressed by it; see Habermas 2006); at a smaller level, however, it may be possible to understand networked discussion on specific topics in the blogosphere to constitute what may be described as a public spherule (Bruns, 2008). 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. (For a full outline of the methodology, see Bruns *et al.*, 2008.)

Research framework

To establish a solid quantitative picture of blog-based topical discussion networks and their cluster patterns, 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:



David Hicks blog posts (Technorati: some authority)

Map Details:

Author: Axel Bruns
Email: a.bruns@qut.edu.au
Crawl start: 1 Mar 2007 - 23:23
Crawl end: 2 Mar 2007 - 12:42
Privilege starting points: off
Analysis Mode: page
Iterations: 2
Depth: 2
Node count: 98

Map generated from Issuecrawler.net by the Govcom.org Foundation, Amsterdam.

Legend:

● (.com) ● (.org) ● (.com.au) ● (.net) ● (.co.uk) ● (.edu) ● (.info)

Select links layers:

links on ▽

Statistics:

○ **roadtosurfdom.com**

Destination URL: http://www.roadtosurfdom.com/
Page date stamp: none
Links received from crawled population: 4648

Links from network (1 - 20)

1. landownunder.blogspot.com
2. willtypeforfood.blogspot.com
3. wsacaucus.org
4. venividiblogi.blogspot.com
5. wbbaboon.com
6. rankandvile.dailyflute.com
7. rwdb.blogspot.com
8. scienceblogs.com
9. southsearepublic.org
10. stoush.net
11. timblair.net
12. polemica.info
13. johnquiggin.com
14. kevgillett.net
15. barista.media2.org
16. benambra.org
17. bunyip.blogspot.com
18. castironbalcony.media2.org
19. chasemeladies.blogspot.com
20. crazybrave.net

Links to network: 42 < | >

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

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:

1) 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.

2) 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).

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*.

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,

1) Patterns of interlinkage between contemporaneous blog posts indicate the existence of a network of *debate* on specific topics. Posts with many incoming links may make an important (possibly controversial) *original* contribution; posts with many incoming and outgoing links may make an important *discursive* contribution; posts with many outgoing links may be *introductions to* or *summaries of* ongoing debate.

2) 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 *hubs* in the network; sites with many incoming links may be central *sources* for information; sites with many outgoing links may be *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 have significant authority overall, while sites whose centrality is high only for specific topics have significant authority only for those topics.

Research methodology

The three key elements of the research process are data gathering and processing, content analysis, and link network analysis. Subsequently, it is also possible to extract and identify common patterns and interrelations between content and 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

Most blogs offer RSS feeds which alert subscribers to new posts, but RSS feeds in themselves are an insufficient data source: some contain only excerpts from whole posts, and many do not contain links, images, or other functional elements of the blog posts. 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. This means that scraped blog pages must be further processed in order to separate salient content (the blog posts itself) from ancillary material; in the process, other salient elements (blogrolls, comments) can also be extracted 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 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. 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. 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 all blogs.
b. Change of themes over time.
c. Identification of key themes for individual bloggers or groups of blogs.
d. 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 post level.
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. 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 automated methods.

Our presentation at ISEA2008 will demonstrate this research approach in practice, and showcase early findings from an exploratory study of the Australian political blogosphere. (For a full outline of the methodology, see Bruns *et al.*, 2008.)

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).

Bruns, A., J. Wilson, B. Saunders, T. Highfield, L. Kirchhoff, and T. Nicolai. 2008. "Locating the Australian Blogosphere: Towards a New Research Methodology." http://eprints.qut.edu.au/archive/00013427/ (accessed 5 May 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.