



COVER SHEET

This is the author version of article published as:

Bruns, Axel (2007) Methodologies for Mapping the Political Blogosphere: An Exploration Using the IssueCrawler Research Tool. *First Monday* 12(5).

Copyright 2007 Axel Bruns This work is licensed under a Creative Commons Attribution-Noncommercial-Share Alike 2.5 Australia License



Accessed from <http://eprints.qut.edu.au>

Methodologies for Mapping the Political Blogosphere: An Exploration Using the IssueCrawler Research Tool

Dr Axel Bruns
Media & Communication
Creative Industries Faculty
Queensland University of Technology
Brisbane, Australia
a.bruns@qut.edu.au – <http://snurb.info/>

Abstract

This paper explores methodologies for using the IssueCrawler research tool to map the interconnections of individual blogs in sections of the blogosphere. It uses the case of Australian-born Guantanamo detainee David Hicks as a case study, mapping the distributed discussions of this case in that part of the Australian blogosphere which is concerned with debating news and politics. Its findings indicate the presence of a strong and sustained engagement with this case by Australian political bloggers, and point to a tendency for discussions to cluster around a handful of sites which are defined by their political orientation. The network maps also suggest a lack of sustained coverage of the case by bloggers outside of Australia, and indicate only limited engagement between bloggers and the mainstream media.

About the Author

Dr Axel Bruns lectures in the Creative Industries Faculty at Queensland University of Technology in Brisbane, Australia. He is the author of *Gatewatching: Collaborative Online News Production* (New York: Peter Lang, 2005) and the editor of *Uses of Blogs* with Joanne Jacobs (New York: Peter Lang, 2006), and is currently developing *From Production to Prodisage: The Rise of Collaborative Content Creation*, forthcoming from Peter Lang in 2007/8. His book *Gatewatching* was nominated for the Communications Policy Research Award at Fordham University's Donald McGannon Communication Research Center. Axel is General Editor of *M/C – Media and Culture* (<http://www.media-culture.org.au/>); his blog is located at <http://snurb.info/>.

Methodologies for Mapping the Political Blogosphere: An Exploration Using the IssueCrawler Research Tool

The growing literature on political blogging demonstrates the considerable interest of academic researchers in understanding the potential role blogs and bloggers may play in the networked mediasphere (see e.g. Bruns & Jacobs, eds., 2006, and here esp. Bahnisch, 2006). Indeed, a number of cases (mainly from U.S. politics) have now emerged which saw political bloggers play a crucial role in public events – this includes the demise of Republican leader Trent Lott (Shachtman 2002), the ‘Rathergate’ affair around falsified documents of George W. Bush’s National Guard service records (Cornfield *et al.* 2005), and the eventual media coverage of images of soldiers’ coffins returning from Iraq, which began only after one blogger’s request for access to such images under freedom of information legislation (Benkler, 2006). (Australia has yet to see its bloggers have a similarly visible impact.)

In spite of such anecdotal evidence, as well as the demographic data beginning to be gathered by the Pew Institute’s Internet & American Life Project and other bodies (see e.g. Rainie, 2005), the exact make-up of and flow of information across the networks that are the domestic and international blogospheres is as yet understood only to a limited extent. Recent research has augmented content analyses, ethnographic research, and studies of specific events in the blogosphere, by adding a new approach: the analysis of automated ‘network crawls’ which from a given starting set of Web pages iteratively follow the available hyperlinks and analyse patterns of interconnection across the population of Websites discovered in the process. (This approach is similar to that used on a much larger scale by the crawlers of search engines as they discover and analyse the content of the entire World Wide Web.)

A key tool for this approach to studying the blogosphere (as well as any other collection of networked Websites, of course) is IssueCrawler, a publicly available crawler system offered by the Amsterdam-based Govcom Foundation at www.issuecrawler.net. In principle, IssueCrawler operates much like the mainframe computers of the Internet’s early history: researchers using the tool share its processing power and queue up their crawls which are processed two at a time, and (also in the spirit of transparently sharing the results of academic research) the datasets produced from their efforts are publicly available to all registered users of the site.

A wide range of studies of Web-based networks using the IssueCrawler tool has already been published (see e.g. Siapera, 2006, or McNally, 2005 for particularly interesting analyses), but these continue to focus in the main on American or international issues. This paper both addresses a number of the methodological questions related to using IssueCrawler to specifically map networks of Australian blogs (as opposed to other, perhaps more static interlinkages of generic Websites at an international level), and presents the results of some initial research into the Australian political blogosphere’s ongoing coverage of the debate surrounding the last remaining Australian-born Guantanamo Bay detainee, David Hicks, who has been in U.S. custody without charge since December 2001 (see *Wikipedia*

2007)¹. While necessarily preliminary, this research both acts as proof of concept and a model for further studies of the Australian blogosphere, and provides valuable new insights into the shape of the Australian political blogosphere and its interconnection with its international counterparts and the mainstream media.

A Brief Introduction to IssueCrawler

The overall operation of IssueCrawler is outlined in some detail in the tool's documentation itself (see esp. Govcom 2004). It begins from a set of starting points or seeds: a list of URLs which is selected by the researcher and defines the 'issue' which is going to be examined by the crawler. The seeds will significantly affect the scope and shape of the resultant network: they are equivalent to a set of coordinates around which geographical terrain is to be mapped, and it is therefore incumbent on the researcher to consider the implication of their seed choices (indeed, researchers may need to conduct a number of exploratory crawls to better understand the implications of their choices).

From here, the crawler gathers the links present in the starting Web pages, and then searches the pages which these links point to, to in turn identify their outlinks; depending on crawl settings, it repeats this process up to three times (this is known as the crawl depth). To further narrow down these results to the core network related to the crawl, the crawler also performs what is described as 'co-link analysis': from all links discovered during the crawl it filters out only those which are reciprocal at least to some extent – that is, it identifies sites which are linked to by at least two of the starting points and which can therefore be considered to be at least part of a loosely interconnected network of Websites. Such co-link analysis can be repeated up to three times (through the iterations setting) – thus, where an increase in crawl depth means that a larger neighbourhood of the seeds is explored for its linkage patterns regardless of the quality of its ties to the core network, an increase in iterations means that such exploration takes place only in areas which were already identified as belonging to the network during earlier iterations. It is also possible to privilege the starting points of a crawl in subsequent iterations – with this setting checked, starting points will continue to be included in later iterations even if the first iteration found that these starting points were in fact non-important nodes (such as simple links directories rather than inherently meaningful pages, for example). Additional settings relating to the crawl include various limitation options (setting maxima for various aspects of the crawl in order to ensure that pages or sites with a great number of links, or highly interconnected networks, do not overload the IssueCrawler), as well as a choice between link analysis on a per-site and a per-page basis (allowing a choice between a coarser and a more fine-grained analysis).

¹ In the time since the completion of this article, Hicks has entered into an unusual plea bargain deal, pleading guilty to a lesser charge in return for a sentence which would see him returned to Australia and released after nine months of further imprisonment. Reportedly, the deal also includes a one-year gag order for Hicks, and requires him to renounce earlier statements that he was mistreated in captivity (see e.g. BBC News, 2007).

The latter choice is likely to be of particular importance in the context of studying networks of blogs: it is common practice amongst bloggers to include an extensive 'blogroll' on their sites – a lengthy list of other blogs which are of general interest. Per-site co-link analysis (which reduces all crawled pages to their base URL) is likely to be contaminated quickly by the presence of such blogrolls: their existence makes it likely that the crawler will find a fairly close connection between a very large number of blogs even if the blogroll links are the *only* form of interlinkage between these sites. Per-page analysis, on the other hand, while not entirely eradicating the influence of the blogroll, will nonetheless mitigate against it: here, individual pages on a site are treated individually, which means in particular that a blogroll link to the base URL of another blog is non-identical to a deep link to a specific page on a blog. Such deep links are commonly used when one blog post comments on a blog post existing on a different blog, however, which means that per-page analysis is better able to distinguish these conversational links from the static affiliational links in blogrolls. In other words, the resulting patterns of interlinkage identified through per-page analysis are finer-grained and track links between specific pages beyond the base URL; connections which are usually directly indicative of the deliberately made links to specific blog posts which are common to the distributed discussions across posts and comments that are frequently seen in the blogosphere.

Finally, the networks discovered in the process are detailed in the crawl's results, both through a variety of datasets for further processing, through lists and matrices of site interlinkages, and (perhaps most immediately useful) through graphical representations of the networks discovered through the crawl. Such network maps plot the key nodes in the network and the connections between them, with the relative positioning of nodes on the map indicating how frequently links are exchanged between nodes, and the size of nodes on the map indicating either simply the number of inlinks each site received from the network, or an aggregate of inlinks received *and* outlinks made – this latter measure is known as a site's level of centrality, as it indicates that the node is not simply a source of information for others, but also actively connects to other information sources (in other words, it can be described as a 'good citizen' of the many-to-many network, rather than operating on a one-to-many broadcast basis). Finally, arrows attached to the links drawn between nodes in the map also indicate the predominant patterns of interlinkage – showing which sites are predominantly origins and which sites are mostly destinations for Web traffic. This mode of graphically plotting nodes in the network is especially useful to identify clusters of highly interconnected sites: beyond their membership in the overall issue network, such sites are particularly close to one another for various reasons – often because of matching values and beliefs, stylistic commonalities, or shared communities of participants.

It should be noted in this context that of course IssueCrawler's results make no prediction about *actual* flows of traffic across the sites contained within its networks. The existence of links does not guarantee that visitors will indeed follow them. However, the widespread practice of browsing the Web (that is, following its links) in order to learn more

about specific issues makes it likely that information seekers interested in particular topics are going to explore the sites which are part of these topics' existing issue networks, and that on aggregate their browsing patterns are going to resemble the patterns of interlinkage between key sites. This is all the more likely as the patterns of interlinkage between individual sites are also used by search engines such as *Google* as a key determinant of the order of their search results listings – sites identified as prominent nodes in IssueCrawler networks are therefore also likely to be listed as key sources of information when Google or other search engines are searched on topics of relevance to the network.

Further, for the purposes of an enquiry into Weblog networks it is not only (and perhaps not even predominantly) visitor traffic across sites which is of interest, but also the content creation activities of bloggers and the contributors of blog comments. IssueCrawler networks provide a very direct insight into such activities, as (beyond links embedded in the blogs' static pages and blogrolls) the vast majority of links which will be found through the crawl are topical links included in blog posts and comments. Perhaps more so than in investigations of Website networks outside of the blogosphere, therefore, crawls of blogs can generate a very clear indication of the online information sources (within and beyond the blogosphere itself) that are highlighted both by bloggers themselves and by those of their visitors who add comments to blog posts.

Towards a Methodology for Blog Crawls: A Case Study

A number of key factors must be considered when planning a crawl-based analysis of any one part of the blogosphere. Perhaps the most obvious question concerns the choice of starting points, or seeds. As the name of the tool suggests, IssueCrawler is predominantly designed for identifying 'issue networks', that is, networks of Websites which form around the interlinkage and exchange of information pertaining to specific issues or topics. The researcher's first challenge, therefore, is the selection of a set of Web pages (containing links to other, related pages) connected to the issue at hand. For our present purposes of researching conversational patterns within the blogosphere, it is preferable that these pages themselves be blog posts related to the issue (rather than simply the base URLs of blogs covering a variety of topics, or of other pages merely listing generic sites related to the issue).

While for the study of generic issue networks an auxiliary tool for IssueCrawler, scrapeGoogle, can be used to automate the harvesting of the highest-ranked URLs returned by *Google* for a given topical query, this is of limited use for research which specifically focusses on the blogosphere; in most cases, only few, if any, of the results returned by *Google* will be blog posts. The proof-of-concept study performed for the purposes of this article, by contrast, utilised the leading blog aggregator *Technorati*, which currently tracks over 70 million blogs world-wide. For the present study, which focussed on blog-based discussion on the fate of Australian-born Guantanamo Bay detainee David Hicks, *Technorati's* search function was used in early February 2007 to identify the one hundred

most recent posts containing the phrase “David Hicks” in either title or body of the blog post. It should be noted here that a phrase search (as opposed to a search for any posts containing either “David” or “Hicks”) is likely to produce a smaller number of false positives; while the sample obtained through the phrase search may still have contained a small number of posts related to a different David Hicks, this is far less likely – and due to the nature of the analysis which is performed by IssueCrawler, such blog posts would be either disregarded as unconnected ‘orphans’ in the issue network, or (in the very unlikely case that there were a high volume of blog posts about a different David Hicks) would be likely to lead to a network depiction which showed two unrelated clusters within the resultant issue network.

The use of *Technorati* necessarily introduces recency of postings as a selection criterion for the seed URLs. For the analysis of blog-based conversation, this should be regarded as an obvious choice: such conversations typically take place over the course of a few days or, at most, weeks, and starting from the most recent contributions should therefore enable the IssueCrawler to work backwards to earlier postings in the communicative threads of the distributed discussion, identifying cross-linkages and other network connections. By contrast, starting from the most ‘relevant’ blog posts on a given issue (however such relevance may be measured), regardless of posting date, would be more likely to trace a number of individual threads, but may not uncover a network of connected blogs as such threads may be spread across a broad timespan. The *Technorati* method, in other words, enables a network analysis of (broadly) synchronous conversations in the blogosphere.

Further, *Technorati* also provides a coarse classification of the blogs it tracks into high, medium, and low authority categories; such classifications are again based on the number of other Weblogs which link to such blogs, and therefore follow the same logic as *Google’s* PageRank system and IssueCrawler itself. This classification can be utilised as a useful pre-crawl filter for the set of crawl seeds used, as it enables a selection of seeds which is likely to include more of the major sites in a network – sites which should therefore contain a large number of crawlable links to other network nodes. (This reduces the risk that a more randomly compiled set of starting points is accidentally biased towards a subset of the network. After all, as the main purpose of issue network studies is to identify the key nodes in the network, *starting* from a selection of these key nodes introduces little additional bias, provided that the crawl setting ‘privilege starting points’ is not used.)

The choice which the different authority classifications offer, then, is one of quality against recency: because the most authoritative blogs, as ranked by *Technorati*, are only a small subset of all available blogs, there is a far lower number of posts containing the phrase “David Hicks” than can be found at the level of “some authority” (which includes posts from both high *and* medium authority blogs) or “a little authority” (which contains the two higher levels and a further, yet larger set of blogs). Additionally, it would of course also be possible to apply no filtering at all, choosing to use posts at any authority level including those with no specific authority rating. On 8 February 2007, the one hundred most recent “David Hicks” posts with “a lot of authority” on *Technorati* ranged back over the previous 97 days, the one

hundred most recent “some authority” posts dated back up to five days, and the one hundred most recent posts with “a little authority” covered only the past three days. (The “high authority” group suits the aim of mapping synchronous blog conversations by starting from roughly contemporaneous posts only inadequately, therefore. Like the other two groups, however, it did contain a notable peak of postings in the days before 8 February as a result of renewed political interest in the Hicks case, as is evident from figure 1.)

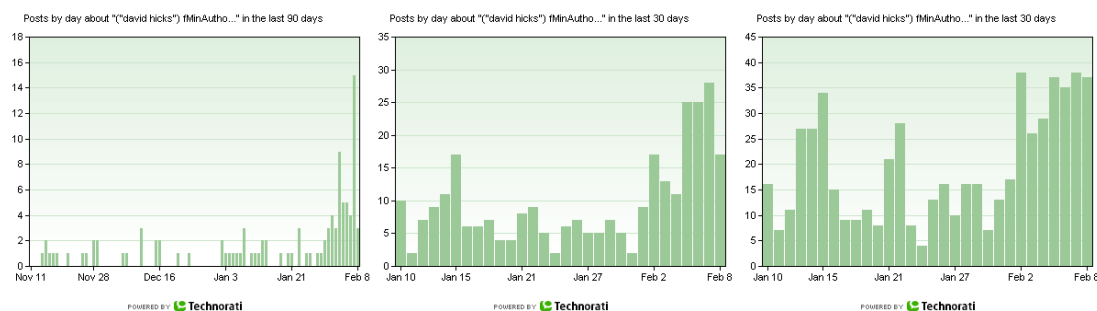


Fig. 1: Traffic volume for “David Hicks” posts with “a lot”, “some” and “a little authority”, dating back from 8 February 2007, as reported by *Technorati*. (The notable increase in traffic in the days before 8 February is due to renewed political debate on the issue since the start of February.)

While overall, *Technorati* is a useful source for a sample of relevant blog posts to be used as the crawl seed, then, the choice of what level of seeds to use must be made with care and according to the intentions of the research project. A number of additional factors must also be considered here: most of all, the highest authority group of blogs may contain only a very limited number of *different* blog sites – in the 8 February sample used here, the one hundred blog posts in that category originated from only 35 different blogs, while one of the blogs, *Road to Surfdom*, was represented 20 times.

Further, due to the way that *Technorati* calculates ‘authority’, this group of blogs is also heavily weighted towards U.S. blogs: authority, after all, is based on a simple numerical measure of how many inlinks a blog receives, without any correction for or limitation to topical or geographic boundaries. Even though their authority in their respective domestic contexts could be considered equivalent, then, a U.S.-based blog which receives inlinks from ten percent of its compatriots would have a significantly higher authority rating than an Australian blog receiving ten percent from *its* compatriots, and a moderately popular blog on a highly popular issue might receive a higher authority rating than a highly regarded blog on a minority issue, simply due to the inevitable difference in the size of each country’s or issue community’s local blogosphere. In other words, unless they receive wide international recognition, fewer Australian blogs are likely to enter the “high authority” category, which is instead set to be dominated by blogs operating in populous nations or large interest communities.

This observation is particularly pronounced in the context of largely domestic Australian topics or for issues which are of relatively limited public appeal. In such contexts, lower-ranked authority groupings on *Technorati* are likely to produce more useful and more

diverse seeds. (Again, it is useful to keep in mind the crawler's way of working in this context, too: one or two Australian blogs in an otherwise U.S.-dominated sample are likely to be sidelined from the resulting network graph if their level of interconnection with American sites in the sample is not found to be on par with the interconnection amongst the U.S. sites, even though their very presence in the sample designates them as the tips of an otherwise unseen iceberg of less authoritative domestic Australian blogs.)

Additionally, as noted above, especially for studies which are interested in creating a snapshot of the shape of a blog network at a given moment, seeds which cover a large date range (in the present case, over three months at the "high authority" level) may be counterproductive. On the other hand, seeds which are based on posts from an overly narrow date range may be subject to random variations beyond what the researchers have intended – seeds with posting dates ranging in hours rather than days, for example, might be biased towards whichever side of the globe had been awake in the hours before the sample was created. (One approach to addressing this problem would be to increase the number of seeds beyond the one hundred most recent posts, until they do cover a date range that is deemed sufficient.) By this reasoning, then, each of the sets of the one hundred most recent posts listed at a medium and low authority on 8 February 2007, covering five and three days respectively, would seem to be useful starting points for a study of the blogosphere's reaction to a current, fast-moving issue; the "high authority" seed should be considered less suitable.

Finally, the practicalities of discovering posts relevant to a specific topic through a *Technorati* search must also be considered. For some topics, no one obvious and distinct phrase might be available – blog posts discussing the upcoming federal elections in Australia might include any or none of the terms "federal election", "John Howard", "Kevin Rudd", "Liberal Party", "Labor Party", etc. Other topical terms, such as "global warming" (and indeed, "Liberal Party" or "federal election"), may be too broad and generic to produce nationally specific results. It may be useful to combine seed sets discovered through searches on a number of relevant terms, or to manually filter search results for Australian-based blogs only. The present case study remains relatively unaffected by such issues: "David Hicks" is a distinct phrase which can reasonably be expected to be present in virtually any blog post discussing the case, and (as the very limited blog post traffic at the "high authority" level also indicates) does not appear very frequently in blog posts originating from outside Australia.

Crawl Results and Discussion

In order to test the methodology described above, and to compare the results obtained from seeding otherwise identical crawls with sets of the one hundred most recent postings at "a lot of", "some", and "a little authority" as listed by *Technorati*, then, three crawls originating from blog posts containing the phrase "David Hicks" were enqueued in IssueCrawler. Crawl depth and number of iterations were left at the IssueCrawler default setting (two for both), as were other settings. Due to some intermittent technical difficulties with the tool, these ran between

28 February and 3 March; this meant, of course, that in addition to the extant postings up to 8 February they could also be expected to discover additional blog posts and other relevant Web pages added to the network since that time. Analysis of the crawl results in this article will focus predominantly on the graphical network maps produced by IssueCrawler; additional and possibly more in-depth information can be obtained from the detailed crawl datasets themselves, but such deeper analysis is beyond the scope of this paper and may be discussed in another publication. Interested readers can also directly access the crawl results in various formats on the IssueCrawler Website, of course; the relevant crawls are accessible on the IssueCrawler site as #307617 (“some authority”), #307618 (“a lot of authority”), and #307619 (“a little authority”). Further, the interactive SVG versions of the network graphs presented in this paper can also be accessed online (Bruns, 2007).

It is immediately obvious that the three crawls produced three very distinct and divergent depictions of the issue network around the David Hicks case. This is a necessary result from the three different sets of seeds, and much of the analysis which follows will focus on these differences. What will emerge very soon is that the three network depictions are complementary to one another rather than contradictory: they can be seen to map different layers, or – perhaps more in keeping with the map metaphor – different scales of magnification of the Australian and international political blogosphere map (or more precisely, that part of the map which pertains to the David Hicks case).

High Authority Crawl

The maps produced through the crawl seeded with blog posts from sites identified by *Technorati* as having “a lot of authority” (figs. 2, 3) show a relatively impoverished network. While the map in figure 2 includes a number of well-known (and mainly U.S.-based) political blogs, whose importance is confirmed through the node size which here shows the number of inlinks received from the overall network identified through the crawl, the relatively uniform geometrical separation of the nodes shows that there is no close interrelation between these nodes in the present context – no sustained discussion of the Hicks case appears to take place across the network existing between these nodes. Further, a number of mainstream news Websites (*New York Times*, *Los Angeles Times*, *Fox News*, *CNN*, *The Guardian*) appear throughout this graph, suggesting that the network graph is mainly representative of an informational rather than a conversational network (the shape which would be expected for a fully developed blog network); even the pundit-bloggers seen here may be included in their role as pundits (that is, sources of controversial opinion) rather than as bloggers (conversants in a distributed discussion).

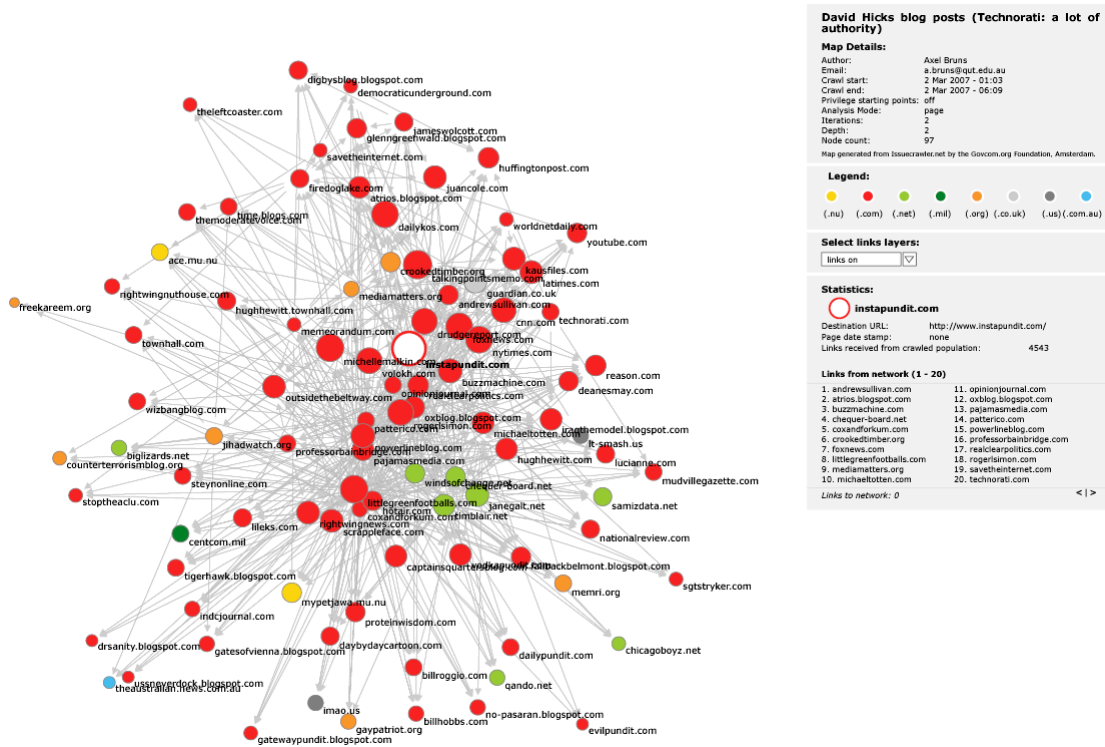


Fig. 2: Network resulting from high-authority seeds; relative node size determined by number of inlinks.

This observation becomes even more obvious through the map in figure 3, which plots node size not simply according to the number of inlinks received, but according to the aggregate of inlinks and outlinks for any given site (their 'centrality' to the network). Few of the sites seen here (including the usual suspects, U.S. pundit-bloggers *Instapundit* and *Talking Points Memo*, as well as the commercial conservative group blog *Pajamas Media*) reach a significant node size on this map, which indicates that at least in relation to the David Hicks case they do not participate in a distributed back-and-forth conversation across the blogosphere. There is also an almost complete absence of Australian blogs or other Australian sites in either of these maps, except for the Website of *The Australian* newspaper (which appears here mainly as an additional but relatively unimportant source of news reports that U.S. bloggers link to, but which does not link back to the network), and the site of right-wing Australian political blogger Tim Blair, who receives a sizeable number of inlinks from politically aligned U.S. bloggers like *Pajamas Media*, *Little Green Footballs*, or *Instapundit's* Glenn Reynolds.

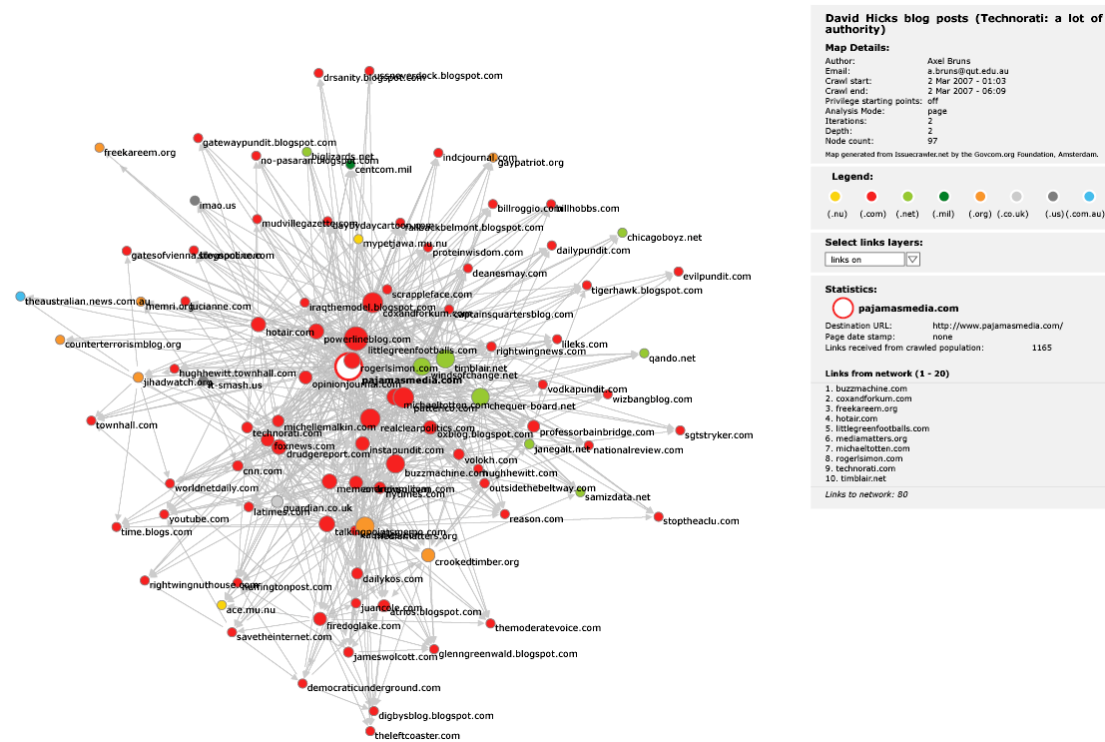


Fig. 3: Network resulting from high-authority seeds; relative node size determined by site centrality.

On the evidence available from this “high authority” crawl, then, at this level there is no strong network of bloggers and their postings discussing the Hicks case. While figure 2 does indicate a set of moderately well interlinked sites, it is likely that these sites appear connected to one another not because they are each interested in discussing David Hicks, but rather due to their participation in other, more immediately engaged distributed conversations across the blogosphere on domestic and international political issues in the United States. The very limited network visible for the David Hicks issue, then, might be assumed to constitute merely a “data shadow” of the actual U.S. political blogosphere. (This interpretation is further strengthened if we take into account the 97-day spread of seeds for this crawl – had the set of seeds been limited to a shorter time frame, it is likely that even more nodes would have disappeared from the network depicted in figs. 2 and 3.)

This points to an important fact to keep in mind when analysing IssueCrawler data: crawls which are set to use a high number of iterations or high level of crawl depth may proceed beyond what can be regarded as the immediate issue network under investigation, if the neighbourhood of the initial seeds does not immediately constitute a strong network in itself. Because of the ephemerality and permeability of Website networks (especially amongst blogs) and the interlinkage of sites across issue networks, IssueCrawler crawls may occasionally ‘jump’ issue networks, in other words. Further, this may be particularly easily possible in cases similar to that shown in figs. 2 and 3, where secondary, perhaps more strongly formed networks are easily found by the crawler and prove more apparently fruitful objects of analysis.

Finally, it is also worth noting that in spite of being represented amongst the one hundred “high authority” seeds with no less than twenty different post URLs, the Australian-based blog *Road to Surfdom* does not appear in either of figs. 2 and 3. This confirms the assumption stated above that even in spite of such strong authority ratings, individual Australian blogs which are not well connected with their international peers will ultimately be treated as network orphans by IssueCrawler, and will therefore be excluded from the resultant network graphs. Beyond the present case study, this would seem to indicate that a limitation to “high authority” posts for the collection of seeds is usually not advisable as it may exclude important nodes either immediately, or indirectly through exclusion effects at the crawler stage. (The only exception from this rule would exist in cases where the bulk of the blog-based discussion on an issue can confidently be assumed to be carried out in, or at least closely linked to, “high authority” blogs.) It may also be possible to mitigate against the cut-off effects of using “high authority” seeds by using more inclusive crawl settings (a greater crawl depth, for example); this will need to be confirmed through further study.

Medium Authority Crawl

The network appears markedly different for the second crawl, which used the one hundred most recent blog posts of “some authority” (that is, medium or better) as a starting point. As noted, the seeds for this crawl contained a far more recent sample of blog posts (stretching over five days rather than over the 97 days covered by the “high authority” sample), and also contained a significantly larger proportion of Australian-based blogs. Largely due to this difference, the resultant network graphs now show clear evidence of a well-formed network (regardless of whether node size is determined by inlinks received or by the node’s centrality to the network). Further, the network also exhibits a complex structure including a number of distinct clusters and a periphery of nodes which appear to be fairly closely connected to one another but remain distant from the core of the issue network.

Most importantly, however, the network centre is very strongly biased towards Australian blogs and other Websites (as would be expected for a political issue that is of interest mainly to Australian contributors). Blogs such as *Larvatus Prodeo*, *Road to Surfdom*, and *John Quiggin*, which on empirical evidence would be believed to be central to issue networks in Australian politics, do indeed feature prominently, while other well-known Websites including *Tim Blair*, *On Line Opinion*, *Crikey*, and the NewsCorp columnist ‘blogs’ at *Blogs.News.com.au* are also visible in the network.

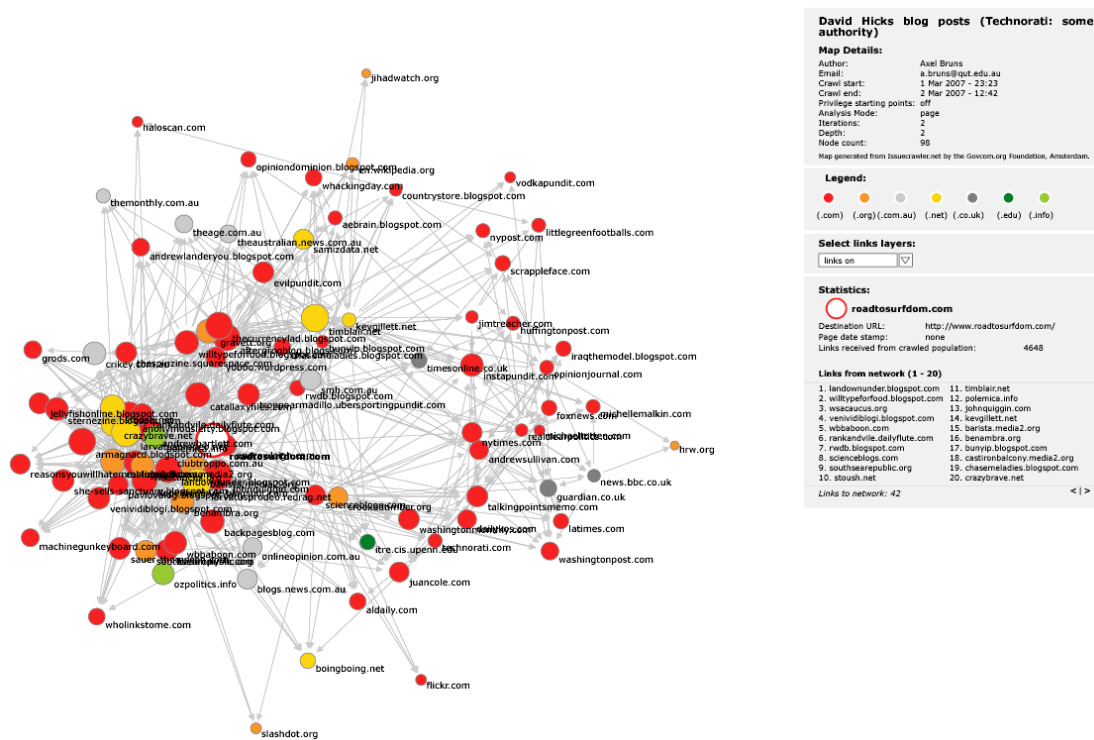


Fig. 4: Network resulting from medium-authority seeds; relative node size determined by number of inlinks.

The tendencies towards clustering which are evident in figure 4 deserve further attention, however. The core of the network is clearly located in the bottom left quadrant of the graph, and is centred mainly around what can comfortably be described as left-of-centre political blogs. Given that the major developments in the Hicks case during February focussed on calls for the federal government of Prime Minister John Howard to take a more proactive approach to convincing the U.S. government to begin either trial or repatriation proceedings, such strong activity in camps oppositional to the ruling conservative Coalition is perhaps what should be expected here.

Beyond such obvious findings, the minor clusters in the network are of particular interest. What emerges from the graph is that there are a number of separate sub-groups of blogs which are strongly connected amongst themselves, but link only much more loosely to the core network. One such cluster consists of a dyad of international academic blogs (*Crooked Timber* and *ScienceBlogs*) in the bottom centre of the map, which may have been drawn into the debate through overlaps between international and Australian academic circles, and (in turn) between Australian academics and Australian political activists especially on the left – indeed, it should be noted here that while staffed by academics, both blogs very frequently also address political and social themes.

Another cluster, located on the map above the main cluster, contains a group of blogs including *After Grog Blog*, *The Currency Lad*, *Will Type for Food*, *Gravett*, *Yobbo*, *Tropo Armadillo*, and *Catallaxy*, and is somewhat less clearly explicable; three of the blogs contained here (*The Currency Lad*, *Gravett*, and *Tropo Armadillo*) have temporarily or

permanently ceased publication, while the others continue to post regularly. The topics addressed by these still-active blogs are relatively wide-ranging (they are not limited to political issues only), and so they could be considered perhaps as a second division of occasionally political Australian blogs: interlinked amongst themselves yet at a distance from the centre of Australian political blogging. The inactive blogs amongst this group may be included here because of their participation in the cluster in the past (as the IssueCrawler may have encountered it in archival posts), or because links in outdated blogrolls – sidebars listing fellow bloggers which are a common feature of blog Websites – still continue to point to these defunct sites. This, then, would count as evidence that the crawl did extend beyond a mapping only of the immediately recent discussion of the Hicks case; such ‘timeframe creep’ may be avoided by setting tighter crawl limits for example through the crawl depth or iteration levels.

There is also a more loosely structured cluster around *Tim Blair*, *Kev Gillett*, *RWDB*, *Bunyip* (another discontinued site), and also including *Chase Me Ladies* (one of the few non-Australian blogs included in the clusters identified here), as well as the *Sydney Morning Herald* newspaper’s Website – the latter being included here most likely only because of its role as a key source of news stories for this cluster (closer analysis of the crawl data shows that *SMH*’s only significant outlink destination is to its Melbournian sister publication, *The Age*). Given the political stance embraced by these blogs, this cluster could perhaps be regarded as the right-of-centre counterpart to the more strongly developed mainly leftist main cluster; this need not be seen as a reflection of the *Sydney Morning Herald*’s political positioning, however: its relative centrality to this cluster could just as well indicate that it was subject to frequent criticism for its coverage of the Hicks case, or that it was merely seen as the most useful news source to link to.

Finally, figure 4 also shows a smattering of international blogs and news sources, mainly on the right side of the map. The placement of these sites on the map shows that they are not central to the core of links exchanges for this issue network (at least as it appears at a “some authority” mapping level); indeed, many such sites receive significant inlink traffic at best from the minor clusters discussed above. It is likely (and the relative affinity of some such peripheral sites on this map may point to this) that a number of these sites *are* part of issue clusters on other topics (such as topics more closely related to domestic U.S. politics, for example); if so, this group of peripheral sites hints at the existence of such clusters much in the same way that the loose network of U.S.-based political sites identified through the “high authority” crawl did. Conversely, then, the “some authority” crawl of blog-based coverage of the David Hicks issue could be regarded as having ‘zoomed in’ from the international and U.S. level to the domestic Australian level.

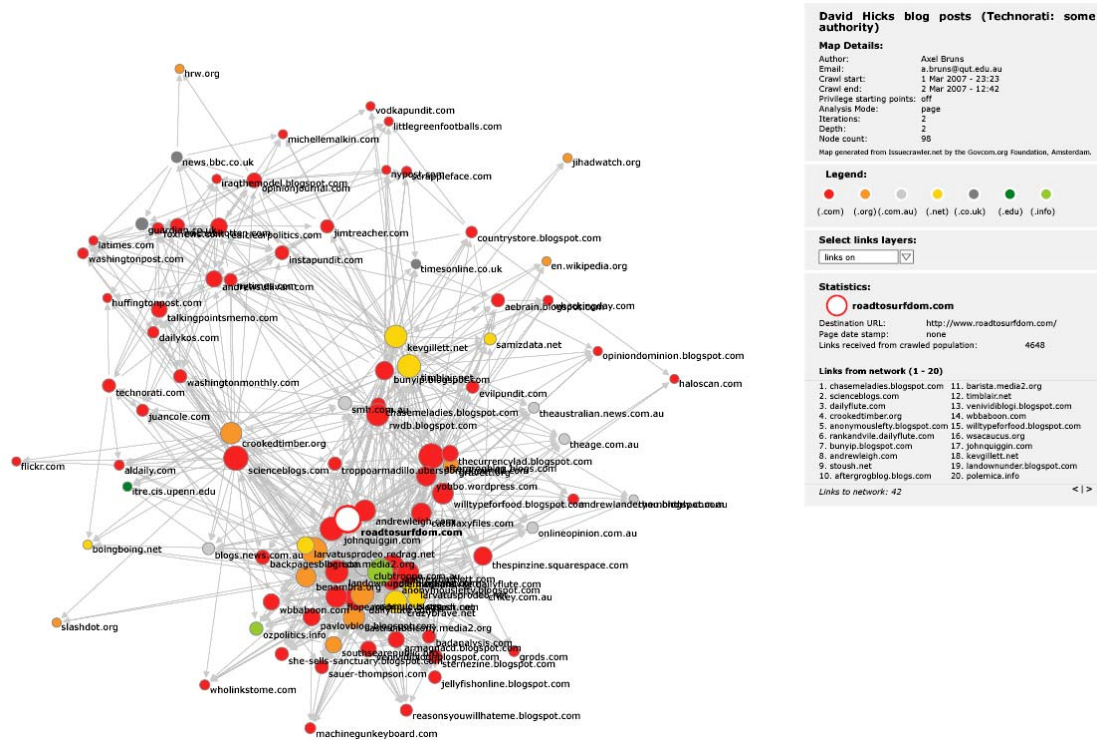


Fig. 5: Network resulting from medium-authority seeds; relative node size determined by site centrality.

The observations above are further strengthened when the “some authority” crawl results are plotted with a focus on site centrality (combining in- and outlinks) rather than merely the sites’ inlinks only. Located here at the bottom centre of the map, the main cluster remains clearly visible, but it is now possible to discern more clearly those sites most centrally involved in a mutual exchange of links (that is, in a distributed discussion on the David Hicks case), as opposed to sites which are talked about but do not talk back, or sites which speak but are rarely spoken to. Indeed, the main cluster may itself show subdivision at this level: a group including *Road to Surfdom*, *Andrew Leigh*, and *John Quiggin* appears now to separate itself slightly from the main core of this cluster which exists around *Club Troppo*, *Anonymous Lefty*, and *Larvatus Prodeo* (this analysis is further complicated, however, by the fact that key leftist blog *Larvatus Prodeo* exists under both larvatusprodeo.net and larvatusprodeo.redrag.net, with both variations appearing separately here). Further analysis, and further iterations of mapping studies beyond the exploration discussed here, may be able to indicate whether the appearance of this new subcluster indicates a genuine and sustained rift in the left-of-centre Australian political blogosphere, or whether it points merely to a temporary division into separate discussion threads during February/March 2007, involving different sub-sections of the same community, as it is common to most forms of computer-mediated many-to-many communication.

The other, previously identified subclusters also remain clearly visible in figure 5; indeed, more so than the major, leftist cluster they retain their respective shapes almost exactly. This may indicate that in the absence of a larger population of participating sites,

small clusters of blogs tend to be more equitable in their interlinkage (and thus, more balanced in the conduct of their distributed discussions) – being small, in other words, small clusters cannot afford for any one member to dominate the discussion, and therefore for other members to contribute less; larger clusters, on the other hand, can comfortably cope with discussion taking place mainly amongst a number of key participants, and with more peripheral members contributing only occasionally. (In this context, it is important again to note that ‘participants’ does not equate here directly to ‘bloggers’: a blog site such as *Larvatus Prodeo* may count amongst its participants both a group of participating bloggers, and a potentially large number of more or less frequent commenters. The prominence of a small number of sites in larger clusters would therefore also indicate that it is these sites where casual users are most likely to contribute a comment.)

By contrast, the difference between figures 4 and 5 (in which the relative size of nodes could be said to indicate a site’s visibility as opposed to its centrality, respectively) is also particularly marked for the Websites of mainstream news media, many of whom can be found in the network periphery. Nodes such as *Fox News*, *LA Times*, or *Washington Post* are markedly smaller in figure 5, which clearly points to the fact that – as we might expect – these sites serve as sources of information for discussion on the Hicks case in the blogosphere, but do not themselves actively participate in the discussion which ensues.

The same comparison of figures 4 and 5 also supports the frequently voiced criticism by bloggers of ‘blogs’ in the mainstream media: *Blogs.News.com.au*, for one, is notably more visible than it is central. This indicates that at least on the David Hicks issue, NewsCorp ‘bloggers’ continue to act more as op-ed writers broadcasting their views to their readership than as genuine bloggers participating in the distributed exchange of views. They fail to engage with the wider blogosphere by including a significant number of links to the views expressed there; this creates an imbalance between incoming and outgoing links which is responsible for the site’s comparative lack of centrality to any cluster in this debate. (Interestingly, the *Tim Blair* node also shrinks, which could indicate that this site, too, acts and is used by bloggers at least some of the time in the manner of a mainstream news site: as a source of information to be discussed or criticised in a blog post, but which does not engage with its critics all too regularly.)

Overall, then, the “some authority” crawl has shown the existence of a strong network of Australian political blogs discussing the David Hicks case, and pointed to the subdivisions (along political and other lines) within this part of the blogosphere. A mapping with node size based on centrality, as shown in figure 5, clearly points to the key sites within this network at which an interchange of opinions occurs between different bloggers. Similar crawls to be conducted in the future are likely to indicate whether the subdivisions in this network are stable across issues and over time, or whether they point to more temporary shifts of attention and participation across the network of blogs. A repetition of present results would indicate that there is a relatively stable division of blog conversations along party lines, with those on the left and those on the right connecting frequently amongst themselves, but less strongly

across to one another; this would also mirror some of the results of studies conducted in the United States which showed a similar polarisation along traditional divides (Adamic & Glance 2005; Hargittai 2005). (That said, it is important to point out that IssueCrawler maps hyperlinks only; it cannot detect the presence of cross-divide conversations in comments attached to the same blog post, for example, unless such comments include links to the commenter's own blog or to other sites supporting their political views. At the same time, it is common practice for commenters to submit the URL for their blog along with their comment, so that comments of this kind would be likely to have an effect on the IssueCrawler results.)

Further, this crawl also appears to indicate the presence of a much more active level of discussion on the left of the political centre than on the right, as well as the existence of a larger number of progressive than conservative blogs; whether this is related to the specific issue at hand (focussing at present strongly on the left's discussion of the fate of David Hicks in the presence of the conservative Australian Federal Government's continued indifference to the issue), or a general feature of the Australian political blogosphere, remains to be substantiated through further study. Similarly, the relative isolation of the domestic blog-based exchange on this topic from the international blogosphere may or may not be indicative of a general disconnection of Australian political blogs from their international counterparts; further studies which focus on issues of strong interest in both domestic and international politics will provide useful points of comparison on this question.

Low Authority Crawl

Finally, then, the results of a crawl conducted from a set of seeds with "a little authority" or better could be seen as further continuing the 'zoom in' to the domestic political blogosphere which marked the difference between high and medium authority crawls. The "low authority" map in figure 6 shows what at first glance appears to be one major cluster of predominantly domestic Australian blogs, with a halo of international sites from *BBC News* to *Amnesty International* and *Flickr* on the periphery as sources of information linked into the distributed discussion of the blogosphere. This cluster covers both the major left-of-centre cluster as well as the subclusters around academic blogs, *After Grog Blog* and others, and *Tim Blair* and the right-wing blogs, which the "medium authority" crawl had discovered.

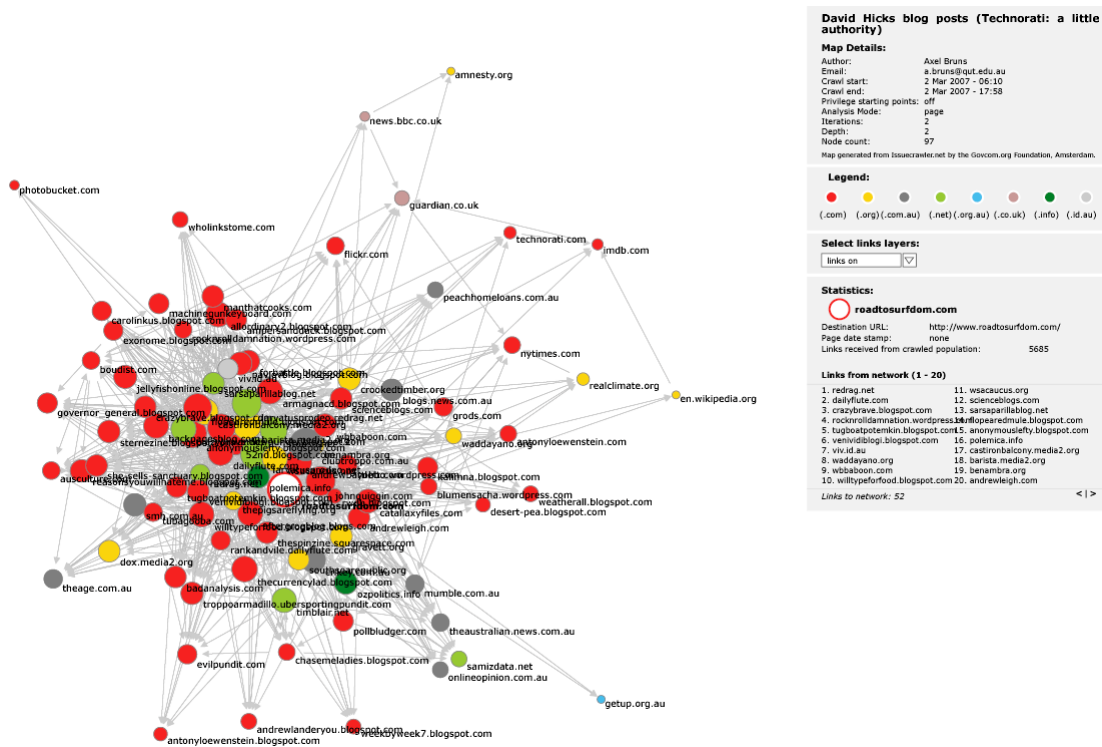


Fig. 6: Network resulting from low-authority seeds; relative node size determined by number of inlinks.

While the relatively unstructured nature of this large cluster appears to run counter to the results of the “medium authority” crawl, plotting the node size according to centrality rather than simply to the number of inlinks received once again helps clarify the picture. When centrality rather than mere visibility is once again taken into account, the map shows a number of subclusters: one at the centre of the map, around blogs such as *Road to Surfdom*, *Polemica / WSA Caucus* (the site exists under both URLs, and both are shown on the map), *Andrew Bartlett*, *Barista*, and *John Quiggin*; one below this around *Daily Flute*, *Flop Eared Mule*, *Crazy Brave*, and others; and a dyad of blogs in *Andrew Leigh* and *Tim Blair*.

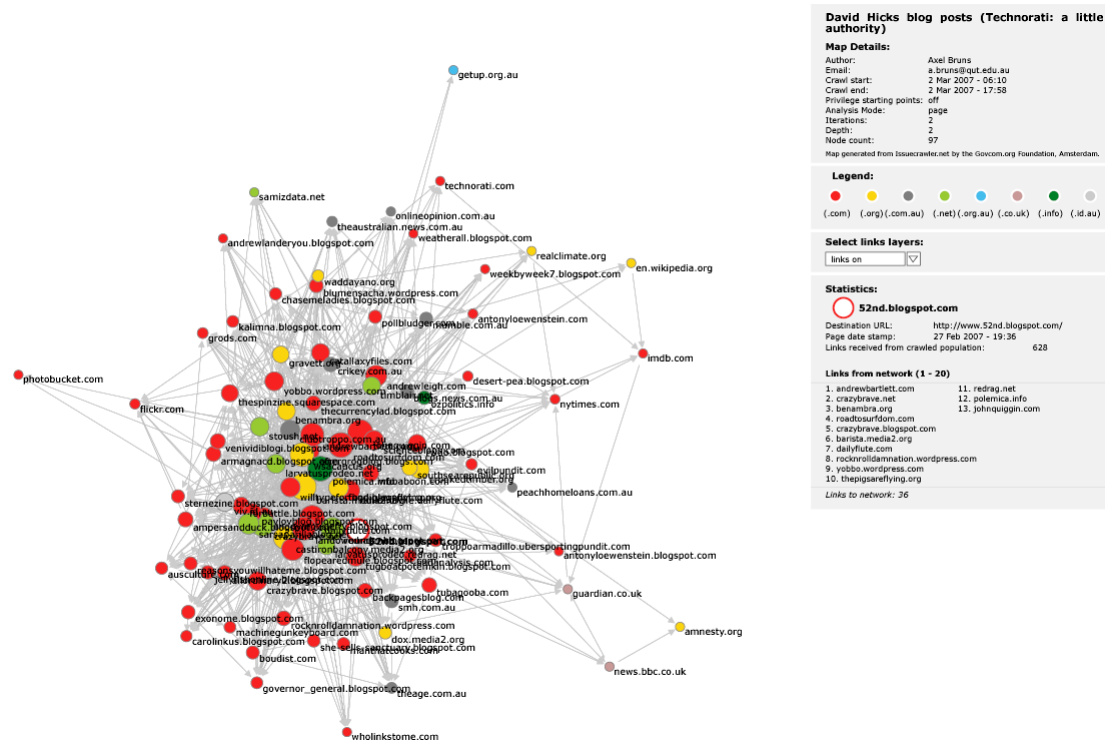


Fig. 7: Network resulting from medium-authority seeds; relative node size determined by site centrality.

The fact that some such blogs were hardly present (if at all) on maps for higher authority levels, and that the placement of individual sites in specific clusters appears to have shifted with the transition from “medium” to “low” authority may be counterintuitive at first; however, IssueCrawler’s underlying process of operation as outlined earlier must again be kept in mind. IssueCrawler maps an issue network beginning from the initial seed sites with which it is provided, and (depending on the settings for crawl depth and iterations) proceeds more or less far beyond the initial seeds; in the first place, however, it maps the network as it is visible from the vantage point described by the seeds. A crawl starting from “medium authority” seeds will therefore be biased towards mapping networks existing at that layer of the blogosphere (consisting of sites with “some authority” or better), and against mapping lower-authority networks. Additionally, it will be more likely to overlook those links between parts of the network which rely on lower-authority sites as intermediaries; a medium-authority crawl may not see such sites, and the links they provide, and therefore regard as unrelated clusters those parts of the network which are not connected through sites of medium authority or better.

Similarly, the predominant absence of non-Australian blogs in this network graph should not be understood to mean that no such blogs take part in debating the David Hicks case. Instead, the “low authority” graph inversely mirrors the “high authority” graph (which contained virtually no Australian blogs), for much the same reasons as discussed above: a crawl starting from a predominantly Australian-based sample of seeds is more likely to treat poorly linked-to international blogs as aberrations to be excluded from the network, just as the

“high authority” crawl was likely to exclude Australian blogs unless (as may be the case for the *Tim Blair* site, which is present at all three levels) a sufficient number of the initial seeds or the sites immediately surrounding them directly link to such blogs.

This would well explain the differences in network structure between the “medium” and “low authority” crawls. It would also suggest that the subcluster around *Daily Flute* and others in the latter crawl constitutes a cluster made up mainly of “low authority” blogs, by contrast with the higher-authority group around *Road to Surfdom* and other sites; the *Daily Flute* cluster could therefore be seen as a kind of ‘minor league’ of blog conversation which does not show up in the “medium authority” map. Additionally, some of the repositioning of the other clusters discovered at “medium authority” now also makes sense: *Tim Blair* and others, as well as the academic blogs subcluster around *Crooked Timber* and *Science Blogs*, now appear as less isolated from the core cluster, because the “low authority” crawl takes into consideration more of the minor blogs which serve to connect them with the main cluster. These blog subclusters may still not connect directly with the core, but they are now more tightly linked at least on an indirect basis. (For the right-wing blogs, this would again mirror studies of U.S. political blogs which saw the connections across political divides as being more weakly formed than those within either camp.)

It should be noted that none of the maps emerging from the three crawls conducted for the present case study is inherently ‘better’ or more valuable than another. Overall, the picture that emerges through this crawl exercise is one of several layers of networked interaction in the blogosphere, which must be examined and understood in relation to one another. Overall (and accounting for the methodological limitations existing at the “high authority” level which we have already discussed), the three maps can be regarded as providing three different levels of ‘zoom’, showing various levels of detail for the blogosphere’s engagement with the David Hicks case: a limited, peripheral coverage of the story at the international level; an ongoing, in-depth discussion of the case by key domestic blogs falling into a range of political and other camps, engaging with a variety of international sites; and indeed beyond this an even more lively range of exchanges within and across these clusters amongst domestic Australian blogs, by bloggers at varying levels of renown and credibility.

A further extension of this study could see such results expanded (for example through crawls using a different choice of depth and iteration settings to produce an even more detailed network map, while avoiding a crossover into topics beyond the Hicks case) and combined (by correlating and overlaying the various maps produced here). It would also be interesting to integrate *Technorati*’s “authority” score directly into the maps – this would produce a three-dimensional map where in addition to the geometric separation of blogs (indicating clustering) and the size of nodes (indicating visibility or centrality) the relative authority of nodes could be plotted for example as peaks and troughs in the network. One question to be answered in this process would be whether those nodes with most visibility or centrality are also generally those which can boast the greatest level of authority.

Of course the underlying research questions will also determine which of the maps produced through such processes will be the most useful. In the context of the David Hicks case, for example, it appears to be relatively clear now that the majority of blogs paying close attention to the case are domestic Australian blogs, and that of these blogs a majority situate themselves to the left of the political divide. Further, detailed analysis of the available data, and further studies, are required to examine why this is so, what the motivations for these bloggers may be, and what if any impact their actions can be shown to have on both Australian politics and Australian media coverage of the Hicks case. It would also be interesting to conduct further investigations into the interconnections between blog-based coverage of this case and the wider protests against U.S. government's use of Guantanamo Bay as a detention centre of deliberately unclear legal status, however, and into where, how, and to what extent Australian blog-based coverage of David Hicks is visible to international audiences.

(It should also be stressed at this point that the left-of-centre bias found here does not necessarily reflect an overall imbalance in the political views expressed in the Australian blogosphere – the present study focussed on blog coverage of one case at one point in time, and the patterns of participation in discussing other topics, or even discussing the same topic at another stage of the news cycle, may well favour other political groupings. Only a continued and topically broad study of patterns of interaction in the Australian blogosphere can hope to produce a more general picture of its predominant political allegiances and of the network clusters which may exist within or across political camps).

Conclusion and Further Outlook

Beyond the blog-based coverage of issues surrounding the detention of David Hicks, which served as a case study here, the results of the present study indicate that the methodology for investigating networks of interlinkage in the blogosphere which has been outlined here appears to be generally appropriate and effective. Utilisation of *Technorati's* three levels of blog authority as a means of filtering recent blog posts for relevant URLs to be used as crawl seeds has been shown to produce markedly different, but mutually consistent results; differences in the results can be explained from the basis of an understanding of the crawl processes employed by IssueCrawler, and can be operationalised in pursuit of specific research aims.

That said, the variations which can be produced through varying the specific crawl settings should be further investigated to deepen our knowledge of what settings are most appropriate to achieve specific research aims. Indeed, the results presented here suggest that the three authority levels (or indeed four, if "any authority" is also counted) available through *Technorati* provide a mechanism to set the "zoom level" (global / intermediate / domestic) for the resultant network graphs, which if used judiciously by researchers may enable them to focus specifically on particularly relevant aspects of the existing network of

blog postings. In this, the “high authority” level appears to be the least immediately useful, however, unless the issue under consideration generates a level of blog traffic which ensures that even at this level of authority the most recent one hundred posts cover only the space of a few days. Conversely, where “high authority” posts about the issue are too rare to produce a compact set of seeds, this level is likely best ignored by researchers.

Further, it should be kept in mind that, as noted, the present methodology will be better suited to investigating clearly definable issues for which relevant blog posts can be reliably identified – and it is worth repeating that specifically *issue*-based networks, which track patterns of conversation across multi-issue blogs, are likely only to result from a ‘per-page’ crawl setting, which the ‘per-site’ option will tend to produce more generic graphs of the patterns of interlinkage between sites as expressed through static features such as blogrolls.

Indeed, there is a need to further investigate the degree to which, as a result of blogrolls and other static links present on blog pages, IssueCrawler explorations of blog networks exhibit a tendency to ‘jump issues’, that is, to map a generic network of blog interlinkages at the base URL level rather than the specific conversation across blog posts at the deep link level. Such tendencies may be identified through a deeper engagement with the raw crawl data than has been possible within the scope of this preliminary exploration; they may be avoided at least in part through further variation of the crawl parameters. It would also be useful if additional functionality could be added to IssueCrawler to exclude particular components of Webpages (such as navigation menus, blogrolls, and other static content) from the analysis, or – technologically more feasible – to include all base URLs from the analysis (this would mean that *only* deep links are processed by the crawler).

It should also be noted, of course, that the discussion in the present article has necessarily focussed mainly on an analysis of the network *maps* produced by IssueCrawler. Additional and very detailed data for the crawl results are available in other forms, including raw XML data, interlinkage matrices, link counts, and other materials, and such data should not be ignored in the further analysis of links networks. Indeed, some further tools for the examination and depiction of IssueCrawler datasets do already exist, and it is hoped that more such tools are in the process of being developed. While the network maps themselves already constitute a very useful resource in examining issue networks, their interpretation is much enhanced by investigating the data behind the maps.

Overall, however, the present study already documents the value of the IssueCrawler tool to the study of patterns of interaction in the distributed discussions of the international blogosphere, and provides a glimpse of the possibilities for further research. It is necessary now to further investigate the outcomes of the present and of other related crawls accessible through the IssueCrawler Website, as well as to extend this research through conducting further such crawls on a variety of topics. Such work crucially requires researchers to develop an in-depth understanding of the underlying crawler functionality and of the effects of seed and parameter choices; the methodology presented here serves as a useful starting point towards further developments in this field.

Acknowledgements

This work is carried out in conjunction with an Australian Research Council-funded project on citizen journalism. My thanks to Associate Professor Terry Flew, doctoral candidate Debra Adams, and other team members for their feedback on earlier version of this paper

Most of this research was conducted during a sabbatical at the Institute for Communications Studies at the University of Leeds, UK. My thanks to Professor Stephen Coleman and the other ICS staff for their hospitality.

Bonne courage also to the Hicks defence team, regardless of the alleged or actual crimes committed by David Hicks. Through their work, they have shone a spotlight on the failure of the Australian and U.S. governments to ensure adequate judicial processes in this case. Democratic governments distinguish themselves from autocratic ones in part by their adherence to fundamental legal principles such as the presumption of innocence, the right to a fair and speedy trial, *habeas corpus*, and the treatment of prisoners of war as prisoners of war under the Geneva Convention, irrespective of case or defendant. The Hicks case suggests strongly that the Howard and Bush Jr. administrations have departed from such basic principles.

References

- Adamic, Lada, and Natalie Glance, 2005, 4 Mar. "The Political Blogosphere and the 2004 U.S. Election: Divided They Blog, *WWW 2005 2nd Annual Workshop on the Weblogging Ecosystem: Aggregation, Analysis and Dynamics*, Chiba, Japan, 10 May 2005, at <http://www.blogpulse.com/papers/2005/AdamicGlanceBlogWWW.pdf> (accessed 16 Mar. 2007).
- Bahnisch, Mark, 2006. "The Political Uses of Blogs," In: Axel Bruns and Joanne Jacobs (editors). *Uses of Blogs*. New York: Peter Lang, pp. 139-149.
- BBC News, 2007, 30 Mar. "Hicks Gets Guantanamo Plea Deal," at <http://news.bbc.co.uk/2/hi/americas/6510899.stm> (accessed 11 Apr. 2007).
- Benkler, Yochai, 2006. *The Wealth of Networks: How Social Production Transforms Markets and Freedom*. New Haven: Yale University Press.
- Bruns, Axel, and Joanne Jacobs (editors), 2006. *Uses of Blogs*. New York: Peter Lang.
- Bruns, Axel, 2007, 18 Mar. "IssueCrawler Results: David Hicks-Related Blog Posts, March 2007," *Snurblog*, at <http://snurb.info/david-hicks-crawl-2007-03> (accessed 18 Mar. 2007).
- Cornfield, Michael, Jonathan Carson, Alison Kalis, and Emily Simon, 2005, 23 May. "Buzz, Blogs, and Beyond: The Internet and the National Discourse in the Fall of 2004," *Pew*

- Internet & American Life Project*, at http://www.pewinternet.org/ppt/BUZZ_BLOGS__BEYOND_Final05-16-05.pdf (accessed 15 Mar. 2007).
- Govcom Foundation, 2004, Sep. "Scenarios of Use for NGOs and Other Researchers," at http://www.govcom.org/scenarios_use.htm (accessed 15 Mar. 2007).
- Hargittai, Eszter, 2005, 25 May. "Cross-Ideological Conversations among Bloggers," *Crooked Timber*, at <http://crookedtimber.org/2005/05/25/cross-ideological-conversations-among-bloggers/> (accessed 16 Mar. 2007).
- McNally, Ruth, 2005. "Sociomics! Using the IssueCrawler to Map, Monitor and Engage with the Global Proteomics Research Network," *Proteomics*, volume 5, pp. 3010-3016.
- Rainie, Lee, 2005, 2 Jan. "The State of Blogging," *Pew Internet & American Life Project*, at http://www.pewinternet.org/pdfs/PIP_blogging_data.pdf (accessed 5 Jan. 2005).
- Shachtman, Noah, 2002, 23 Dec. "Blogs Make the Headlines," *Wired News*, at <http://www.wired.com/news/culture/0,1284,56978,00.html> (accessed 15 Mar. 2007).
- Siapera, Eugenia, 2006. "Multiculturalism, Progressive Politics and British Islam Online," *International Journal of Media and Cultural Politics*, volume 2, number 3, pp. 331-346.
- Wikipedia*, 2007, 16 Mar. "David Hicks," *Wikipedia: The Free Encyclopedia*, at http://en.wikipedia.org/wiki/David_Hicks (accessed 18 Mar. 2007).