The Political Blogosphere and the 2004 U.S. Election: Divided They Blog

Lada Adamic
HP Labs
1501 Page Mill Road
Palo Alto, CA 94304
lada.adamic@hp.com

Natalie Glance
Intelliseek Applied Research Center
5001 Baum Blvd.
Pittsburgh, PA 15217
nglance@intelliseek.com

4 March 2005

Abstract

In this paper, we study the linking patterns and discussion topics of political bloggers. Our aim is to measure the degree of interaction between liberal and conservative blogs, and to uncover any differences in the structure of the two communities. Specifically, we analyze the posts of 40 "A-list" blogs over the period of two months preceding the U.S. Presidential Election of 2004, to study how often they referred to one another and to quantify the overlap in the topics they discussed, both within the liberal and conservative communities, and also across communities. We also study a single day snapshot of over 1,000 political blogs. This snapshot captures blogrolls (the list of links to other blogs frequently found in sidebars), and presents a more static picture of a broader blogosphere. Most significantly, we find differences in the behavior of liberal and conservative blogs, with conservative blogs linking to each other more frequently and in a denser pattern.

1 Introduction

The 2004 U.S. Presidential Election was the first Presidential Election in the United States in which blogging played an important role. Although the term weblog was coined in 1997, it was not until after 9/11 that blogs gained readership and influence in the U.S. The next major trend in political blogging was "warblogging": blogs centered around discussion of the invasion of Iraq by the U.S.[1]

The year 2004 saw a rapid rise in the popularity and proliferation of blogs. According to a report from the Pew Internet & American Life Project published in January 2005, 32 million U.S. citizens now read weblogs. However, 62% of online Americans still do not know what a weblog is.[2] Another report from the same project showed that Americans are turning to the Internet in increasing numbers to stay informed about politics: 63 million in mid-2004 vs. 30 million in March 2000.[3]

A significant fraction of that traffic was directed specifically to blogs, with 9% of Internet users saying they read political blogs "frequently" or "sometimes" during the campaign.[4] Indeed, political blogs showed a large growth in traffic in the months preceding the election.[5]

Recognizing the importance of blogs, several candidates and political parties set up weblogs during the 2004 U.S. Presidential campaign. Notably, Howard Dean's campaign was particularly

[1] http://en.wikipedia.org/wiki/Weblog
[2] http://www.pewinternet.org/PPF/r/144/report_display.asp
[3] http://www.pewinternet.org/PPF/r/141/report_display.asp
[4] http://www.pewinternet.org/pdfs/PIP_blogging_data.pdf
[5] http://techcentralstation.com/011105B.html

successful in harnessing grassroots support using a weblog as a primary mode for publishing dispatches from the candidate to his followers. In the third quarter of 2003, Dean's campaign raised $7.4, of a total $14.8 million, via the Internet, with a remarkably modest average donation of under $100.[6] His fall was as spectacular as his rise, and was likewise fueled by means of the Internet, which served as an echo chamber for the infamous "Dean Scream." Although official campaign blogs played a lesser role in the Bush and Kerry campaigns, with the Bush campaign's blog being criticized as little more than a place to post press releases, both parties launched innovative online campaigns to boost their grass-roots efforts [4].

Signalling the established position of blogs in political discourse, both parties credentialed a number of bloggers to cover their nominating conventions as journalists. This lead to the creation of sites to aggregate the content of these blogs during the national conventions: `www.conventionbloggers.com` for the RNC, and `cyberjournalist.net` for the DNC [7]. Other sites adapted to keep track of an ever proliferating mass of political blogs by creating specialized political blog search engines, aggregate feeds and search analytics, including Feedster (politics.feedster.com), BlogPulse (politics.blogpulse.com) and Technorati (politics.technorati.com).

Weblogs may be read by only a minority of Americans, but their influence extends beyond their readership through their interaction with national mainstream media. During the months preceding the election, there were several cases in which political blogs served to complement mainstream media by either breaking stories first or by fact-checking news stories. For example, bloggers first linked to Swiftvets.com's anti-Kerry video in late July and kept the accusations alive, until late August, when John Kerry responded to their claims, bringing mainstream media coverage.[8] In another example, bloggers questioned CBS News' credibility over the memos purportedly alleging preferential treatment toward President Bush during the Vietnam War. Powerline broke the story on September 9th[9], launching a flurry of discussions across political blogs and beyond. Dan Rather apologized later in the month. A more light-hearted example was the post-presidential debate question "Was Bush Wired?" Salon.com was the first to ask the question on October 8[10], which was then quickly taken up by bloggers such as Wonkette and PoliticalWire.com, and then addressed the next day by the mainstream media.

Because of bloggers' ability to identify and frame breaking news, many mainstream media sources keep a close eye on the best known political blogs. A number of mainstream news sources have started to discuss and even to host blogs. In an online surveying asking editors, reporters, columnists and publishers to each list the "top 3" blogs they read, Drezner and Farrell [5] identified a short list of dominant "A-list" blogs. Just 10 of the most popular blogs accounted for over half the blogs on the journalists' lists. They also found that, besides capturing most of the attention of the mainstream media, the most popular political blogs also get a disproportionate number of links from other blogs. Shirky [14] observed the same effect for blogs in general and Hindman et al. [8] found it to hold for political websites focusing on various issues.

While these previous studies focused on the inequality of citation links for political blogs overall, there has been comparatively little study of subcommunities of political blogs. In the context of political websites, Hindman et al. [8] noted that, for example, those dealing with the issue of abortion, gun control, and the death penalties, contain subcommunities of opposing views. In the case of the pro-choice and pro-life web communities, an earlier study [2] found pro-life websites to be more densely linked than pro-choice ones. In a study of a sample of the blogosphere, Herring et al.[7] discovered densely interlinked (non-political) blog communities focusing on the topics of Catholicism and homeschooling, as well as a core network of A-list blogs, some of them political.

Recently, Welsch [17] studied a single-day snapshot of the network neighborhoods of Atrios, a popular liberal blog, and Instapundit, a popular conservative blog. He found the Instapundit neighborhood to include many more blogs than the Atrios one, and observed no overlap in the URLs

[6] `http://www.gwu.edu/~action/2004/dean/deanfin.html`
[7] `http://www.cyberjournalist.net/news/001461.php`
[8] http://politics.blogpulse.com/04_11_04/politics.html
[9] `http://www.powerlineblog.com/archives/007760.php`
[10] http://www.salon.com/news/feature/2004/10/08/bulge/index_np.html

cited between the two neighborhoods. The lack of overlap in liberal and conservative interests has previously been observed in purchases of political books on Amazon.com [9]. This brings about the question of whether we are witnessing a cyberbalkanization [13, 15] of the Internet, where the proliferation of specialized online news sources allows people with different political leanings to be exposed only to information in agreement with their previously held views. Yale law professor Jack Balkin provides a counter-argument[11] by pointing out that such segregation is unlikely in the blogosphere because bloggers systematically comment on each other, even if only to voice disagreement.

In this paper we address both hypotheses by examining in a systematic way the linking patterns and discussion topics of political bloggers. In doing so, we not only measure the degree of interaction between liberal and conservative blogs, but also uncover differences in the structure of the two communities. Specifically, we analyze the posts of 40 A-list blogs over the period of two months preceding the U.S. Presidential Election of 2004, to study how often they referred to one another and what the overlap was in the things they discussed, both within the liberal and conservative communities, and also across communities. We also study a single day snapshot of over 1,000 political blogs. This snapshot captures blogrolls (the list of links to other blogs frequently found in sidebars), and presents a more static picture of a broader blogosphere.

From both samples we found that liberal and conservative blogs did indeed have different lists of favorite news sources, people, and topics to discuss, although they occasionally overlapped in their discussion of news articles and events. The division between liberals and conservatives was further reflected in the linking pattern between the blogs, with a great majority of the links remaining internal to either liberal or conservative communities. Even more interestingly, we found differences in the behavior of the two communities, with conservative blogs linking to a greater number of blogs and with greater frequency. These differences in linking behavior were not drastic, and we can not speculate how much they correlated, if at all, with the eventual outcome of the election. They were nonetheless interesting, and we believe they show an insightful glimpse into the online political discourse leading up to the election.

2 Methodology

In order to get a representative view of the liberal and conservative blog communities, we cast our nets wide and gathered a single day's snapshot of over a thousand political blogs. Since we also wanted to do a careful study of the heart of the political blogosphere, we then analyzed the posts for the two months preceding the election with a smaller set of 40 influential blogs.

2.1 Calling all political blogs

We set about gathering a large set of political blog URLs by downloading listings of political blogs from several online weblog directories, including eTalkingHead, BlogCatalog, CampaignLine, and Blogarama. The directories had surprisingly little overlap, and occasionally listed conflicting categories for a single blog, even within the same directory. We did not gather the URLs of libertarian, independent, or moderate blogs, which were far fewer in number. We attempted to retrieve a single, 'front' page for each blog on February 8, 2005. From this set of pages, we counted up all citations to weblogs not on our original list. For all weblogs discovered this way that were cited 17 or more times, we then labeled their orientation manually based on posts and blogrolls and added them to the set. We retrieved pages for these additional blogs on February 22, 2005. Neither the directory labels, which often rely on self-reported or automated categorizations, nor our manual labels, are 100% accurate. However, since we are considering the aggregate behavior of well over 1,000 blogs, having a dozen or so mislabeled blogs will not affect our results significantly.

The set we attempted to retrieve initially was surprisingly balanced. There were 1494 blogs in total, 759 liberal and 735 conservative. Of these, we retrieved pages that were at least 8KB in size for 676 liberal and 659 conservative blogs. Some of the blogs which were not retrievable either no

[11] http://balkin.blogspot.com/2004_01_18_balkin_archive.html#107480769112109137

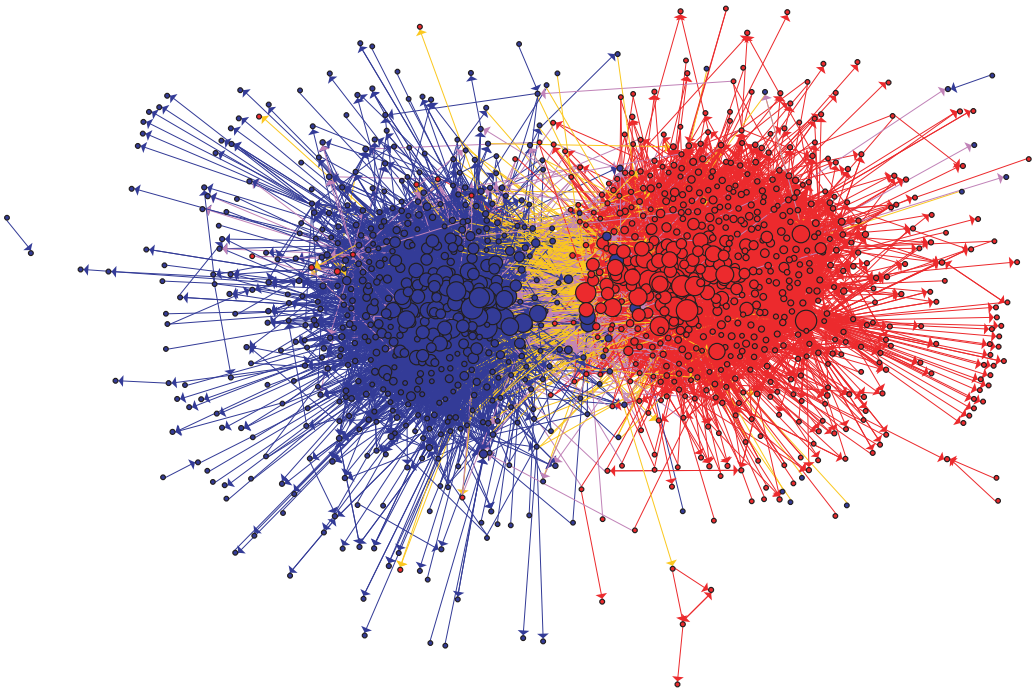

Figure 1: Community structure of political blogs (expanded set), shown using utilizing a GEM layout [11] in the GUESS[3] visualization and analysis tool. The colors reflect political orientation, red for conservative, and blue for liberal. Orange links go from liberal to conservative, and purple ones from conservative to liberal. The size of each blog reflects the number of other blogs that link to it.

longer existed, or had moved to a different location. When looking at the front page of a blog we did not make a distinction between blog references made in blogrolls (blogroll links) from those made in posts (post citations). This had the disadvantage of not differentiating between blogs that were actively mentioned in a post on that day, from blogroll links that remain static over many weeks [10]. Since posts usually contain sparse references to other blogs, and blogrolls usually contain dozens of blogs, we assumed that the network obtained by crawling the front page of each blog would strongly reflect blogroll links. 479 blogs had blogrolls through `blogrolling.com`, while many others simply maintained a list of links to their favorite blogs. We did not include blogrolls placed on a secondary page.

We constructed a citation network by identifying whether a URL present on the page of one blog references another political blog. We called a link found anywhere on a blog's page, a "page link" to distinguish it from a "post citation", a link to another blog that occurs strictly within a post. Figure 1 shows the unmistakable division between the liberal and conservative political (blogo)spheres. In fact, 91% of the links originating within either the conservative or liberal communities stay within that community. An effect that may not be as apparent from the visualization is that even though we started with a balanced set of blogs, conservative blogs show a greater tendency to link. 84% of conservative blogs link to at least one other blog, and 82% receive a link. In contrast, 74% of liberal blogs link to another blog, while only 67% are linked to by another blog. So overall, we see a slightly higher tendency for conservative blogs to link. Liberal blogs linked to 13.6 blogs on average, while conservative blogs linked to an average of 15.1, and this difference is almost entirely due to the higher proportion of liberal blogs with no links at all.

Although liberal blogs may not link as generously on average, the most popular liberal blogs, Daily Kos and Eschaton (atrios.blogspot.com), had 338 and 264 links from our single-day snapshot

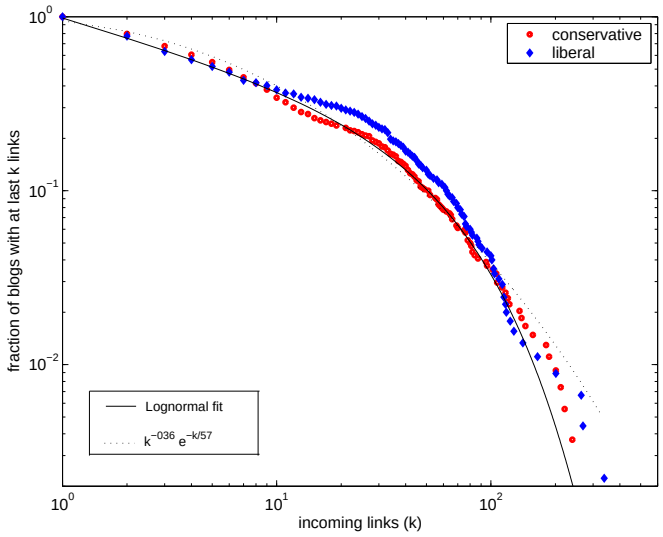

Figure 2: Cumulative distribution of incoming links for political blogs, separated by category. As in [5] we find a lognormal, shown as a dashed line, to be a fairly good fit. A power-law with an exponential cutoff, shown as a solid line, is an even better fit.

of political blogs. This is on par with the 277 links received by the most linked to conservative blog Instapundit. Figure 2 shows that, as is common in nearly every large subset of sites on the web [1, 8, 12], the distribution of inlinks is highly uneven, with a few blogs of either persuasion having over a hundred incoming links, while hundreds of blogs have just one or two.

A small handful of blogs command a significant fraction of the attention, and it is these blogs that we will be analyzing in more detail in the following sections. We will return to a descriptive analysis of the larger set of blogs in 3.5.

2.2 Weblog selection

In order to perform both an in-depth analysis and minimize bias, we decided to work with the top 20 conservative leaning and top 20 liberal leaning blogs, which we identified as follows. First we used page link counts to cull the top 100 or so conservative blogs and the top 100 liberal blogs from the larger list of 1494 political weblogs. Next, we used BlogPulse's index (`www.blogpulse.com/search`) of weblog posts to count up citations to this list of approximately 200 weblogs during the months of October and November 2004. Tables 1 and 2 show the citation count and overall rank among all weblogs for the top 20 lists.

It is interesting to note that the top 20 conservative leaning weblogs fall within the overall top 44 most cited weblogs for this time period. In contrast, the top 20 liberal leaning blogs fall within the overall top 77. This is evidence that bloggers in general link to conservative blogs more than to liberal blogs. Might bloggers "en masse" have predicted the outcome of the election after all?

These lists have some notable omissions. For example, we chose not to include drudgereport.com, which had received 7813 blog post citations during Sep. - Oct. 2004, because of its unusual format and primary function as a mainstream news filter. We also omitted democraticunderground.com, which is principally a message board and secondarily a weblog.

In Tables 1 and 2 we have also included the number of page links from the larger set of liberal and conservative blogs in February of 2005 to the top ranked blogs. The page link counts serve both to validate the popularity of the blogs and to show that the writings of a few bloggers, like Andrew Sullivan and Wonkette have across the board appeal, while most others are read almost exclusively by either the right or the left. It is immediately apparent that page link counts do not produce exactly

overall rank	# post citations	# posts	conservative weblog	# links L-blogs	# links R-blogs
2	8438	924	www.powerlineblog.com	26	195
3	7813	740	instapundit.com	43	234
8	5298	682	www.littlegreenfootballs.com/weblog	10	171
11	3297	66	www.hughhewitt.com	11	146
12	3226	494	www.andrewsullivan.com/index.php	59	86
13	3220	701	www.captainsquartersblog.com/mt	5	117
14	3186	801	www.wizbangblog.com	14	125
16	2781	398	www.indcjournal.com	6	60
17	2773	341	www.michellemalkin.com	10	191
19	2596	1027	blogsforbush.com	4	208
25	2259	87	www.allahpundit.com	2	37
26	2156	100	belmontclub.blogspot.com	3	93
27	1944	50	realclearpolitics.com	13	104
28	1882	633	volokh.com	27	80
35	1570	510	timblair.spleenville.com	7	80
37	1523	428	windsofchange.net	16	65
38	1512	595	www.vodkapundit.com	9	97
40	1468	446	www.rogerlsimon.com	6	74
42	1364	899	www.deanesmay.com	8	79
44	1310	580	mypetjawa.mu.nu	0	51

Table 1: The top 20 conservative blogs by BlogPulse citation count and overall rank (October - November 2004). The two right columns show for comparison how many liberal and conservative blogs from the larger set linked to the blog in February 2005. Also included are the number of posts in our data set for each weblog.

the same rankings as post citations do. In fact, several blogs that are highly ranked according to page link counts, such as scrappleface.com, nationalreview.com/thecorner, thismodernworld.com, and outsidethebeltway.com do not make in into the top 20 lists (just barely missing, however). One factor is that the page links used for the rankings were collected in February 2005, while the post citation counts were tallied in the pre-election months. Even more importantly, as we have already argued, the blogrolls represented in the page link counts may be somewhat stale. Case in point are the the 39 links to allahpundit.com and 23 links to blog.johnkerry.com found in February 2005, even though the former's author had retired in December 2004, and the latter had not been updated since October 2004. On the other hand, weblog rankings via post citation count are sensitive to the time period used for the calculation. The top 20 lists vary from month to month, although the very top most cited weblogs tend to remain the same.

In Table 3 we compare the top 10 blogs derived using BlogPulse data with the ranks assigned by several different ranking sites. TheTruthLaidBear and Technorati rely on link counts, while SiteMeter ranks according to traffic to the blog. Different approaches have different drawbacks. SiteMeter can only rank blogs that use its traffic meter and cannot differentiate unique visitors. Link-reliant rankings are affected by the freshness of the links and run the hazard of being manipulated by inventive bloggers. Despite the differences in ranking algorithms, we observe that the top 10 blogs in each list overlap significantly amongst themselves and with our top 20 rankings.

Because different ranking approaches can produce slightly different sets of top-ranking blogs, we checked that our results are robust with respect to the the particular selection of the top blogs. We ran the set of analysis presented in Section Section 3 with variations in the set of weblogs, replacing some of the lower ranked top 20 with weblogs lower in the list. We found that qualitatively our results remain the same.

overall rank	# post citations	# posts	liberal weblog	# links L-blogs	# links R-blogs
1	10053	1114	dailykos.com	292	46
6	6452	580	www.talkingpointsmemo.com	242	22
7	5468	945	atrios.blogspot.com	230	39
9	4830	502	www.washingtonmonthly.com	165	36
18	2764	409	www.wonkette.com	83	30
24	2277	211	www.juancole.com	149	16
30	1675	550	yglesias.typepad.com/matthew	104	24
33	1621	429	www.crookedtimber.org	81	19
41	1365	348	www.mydd.com	107	8
45	1289	512	www.oliverwillis.com	97	20
48	1268	767	blog.johnkerry.com	21	2
49	1257	607	www.pandagon.net	118	5
54	1191	949	www.talkleft.com	126	15
55	1142	345	digbysblog.blogspot.com	115	3
56	1141	722	www.politicalwire.com	87	16
59	1077	470	www.j-bradford-delong.net/movable_type	98	11
66	1002	722	www.prospect.org/weblog	102	11
68	991	1653	americablog.blogspot.com	64	5
74	947	582	www.theleftcoaster.com	78	4
77	851	115	www.jameswolcott.com	74	6

Table 2: The top 20 liberal blogs by citation count and overall rank according to BlogPulse data (October - November 2004). The two right columns show for comparison how many liberal and conservative blogs from the larger set linked to the blog in February 2005. Also included are the number of posts in our data set for each weblog.

2.3 Data collection

We created a corpus of weblog posts from the top 20 conservative leaning blogs and the top 20 liberal leaning blogs. Table 1 and Table 2 include the number of posts we harvested for each weblog for the time period 8/29/04 - 11/15/04. In all, we collected 12,470 posts from the left leaning set of blogs and 10,414 posts from the right leaning set.

We used BlogPulse's collection system to harvest posts. The strength of BlogPulse's collection system is that it is able to crawl weblog pages and segment the pages into individual posts. BlogPulse currently monitors over 5.5 million weblogs and indexes 450K weblog posts per day. Its coverage falls short of other comprehensive weblog search systems such as Technorati and PubSub because of its requirement that it be able to identify individual full-content posts. Segmentation is a trivial task for a weblog with a full-content feed (assuming that the feed can be automatically discovered). The task is tricker for weblogs with partial content feeds or no feed. In this case, BlogPulse uses a model-based wrapper learner to extract individual posts. Search over this index of weblog posts is publicly available at `http://www.blogpulse.com` [6].

From a corpus of individual posts, it is straightforward to analyze interblog citation behavior, as automatically generated links that appear within the weblog template are removed. What are these automatically generated links? In some blog hosting systems like modblog.com, there are links to recently updated weblogs hosted by the same hosting system included in the template of every blog. Clearly, it would be inaccurate to count these as intentional citations. (However, we did not find examples of this kind of non-intentional linking in our larger sample of political blogs.)

Another set of automatically generated links are blogroll links. In this case, weblog authors clearly intend to link to other weblogs. However, as mentioned earlier, blogrolls tend to grow stale over time if authors do not maintain them actively. Also, it can be argued that linking behavior within posts is more indicative of a blogger's reading activity than are blogroll links.

	Technorati	SiteMeter	TheTruthLaidBear	BlogPulse
1	Instapundit	Daily Kos	Instapundit	Daily Kos
2	Daily Kos	Instapundit	Daily Kos	Power Line
3	Eschaton	Eschaton	Power Line	Instapundit
4	Little Green Footballs	Little Green Footballs	Little Green Footballs	Talking Points Memo
5	Andrew Sullivan	Power Line	Michelle Malkin	Eschaton
6	Wonkette	Wonkette	Talking Points Memo	Little Green Footballs
7	Power Line	Smirking Chimp	Eschaton	Washington Monthly
8	Volokh Conspiracy	Michelle Malkin	Captain's Quarters	Hugh Hewitt
9	Michelle Malkin	Blog for America	Volokh Conspiracy	Andrew Sullivan
10	Lileks	Lileks	Wizbang	Captain's Quarters

Table 3: Different methods produce different rankings, but the overlap in the top rated blogs is high. (Feb. 23, 2005 for Technorati, SiteMeter and TruthLaidBear; October - November 2004 for BlogPulse)

Another advantage of creating a corpus from posts is that, apart from errors in segmentation, there is no duplicate content in the corpus. In contrast, a weblog spider that crawls a snapshot of a weblog at regular intervals, or whenever the weblog updates, will collect snapshots of overlapping content. The resulting data collection would not lend itself to producing accurate analytics.

3 Analysis

3.1 Strength of community

We contrasted the citation behavior in the posts of the top 20 liberal and top 20 conservative blogs. During the two months covered by our analysis, the top 20 liberal bloggers published 12,470 posts, compared to 10,414 for the conservatives. We then counted the number of posts in which each blog cited another blog. If a blog was cited more than once within the same post, the link was not double-counted. We found that liberal blogs cited one another 1511 times, compared to conservatives who cited one another 2110 times. Cross citing accounted for only 15% of the links, with liberals citing conservatives 247 times, and conservatives citing liberals 312 times. The interesting result is that even though the conservatives had 16% fewer posts, they posted 40% more links to one another, linking at a rate of 0.20 links per post, compared to just 0.12 for liberal blogs.

We further found that the citations were concentrated among a smaller subset of the top 20 liberal blogs, but were relatively more distributed among the conservative blogs. Our observations are illustrated in Figure 3. We start out with a connected network of 20 blogs of each kind, with the conservative network having 278 directed internal edges, the liberal one having 218, and 210 cross-edges between them (Fig. 3A). Next we remove any edges that are not sufficiently reciprocated (have fewer than 5 citations in either or both directions). This leaves 40 bidirected edges within the conservative network and 25 for the liberal one, with only 3 reciprocated edges between them (Fig. 3B). Now if we further require that the total number of citations between two blogs be at least 25, then the communities separate completely, and the liberals are left with only 12 edges while the conservatives have 23 (Fig. 3C). Through these visualizations, we see that right-leaning blogs have a denser structure of strong connections than the left, although liberal blogs do have a few exceptionally strong reciprocated connections.

3.2 Varied conversations

The denser linking pattern of conservatives begged the question of whether the conservative bloggers had a more uniform voice than the liberal ones did. We measured this by comparing the pairwise cosine similarity in the URLs that blogs cited in their posts. The cosine similarity is a simple

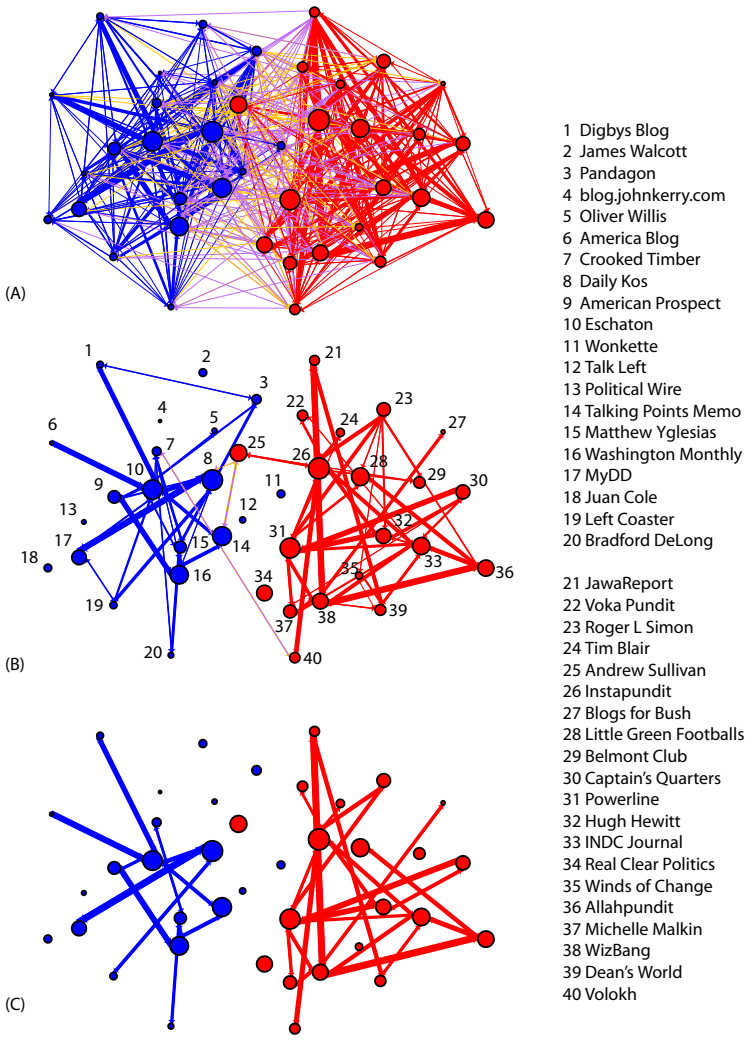

Figure 3: Aggregate citation behavior prior to the 2004 election. Blogs are colored according to political orientation, and the size of the circle reflects how many citations from the top 40 the blog has received. The thickness of the line reflects the number of citations between two blogs. (A) All directed edges are shown. (B) Edges having fewer than 5 citations in either or both directions are removed. (C) Edges having fewer than 25 combined citations are removed.

measure of overlap: $S_{A,B} = \frac{x_A * x_B}{\|x_A\| \|x_B\|}$, where x_A in this case is a binary vector, with each entry being a 1 or 0 corresponding to whether blog A cites a particular URL. We observed that the average similarity between liberal and conservative blogs was quite low, at $S_{avg} = 0.03$. We also found that conservative blogs had a higher similarity on average ($S_{avg} = 0.11$) amongst themselves than did liberal blogs ($S_{avg} = 0.09$). An analysis of variance found this difference to be statistically significant at $p = 0.004$. We found, however, that this difference in similarity was almost entirely accounted for by the conservative blogs' preference for linking to other political blogs. Once we remove from our analysis all URLs pointing to political blogs, the liberal and conservative blogs both had an average similarity of 0.083 and 0.087, a difference that is not statistically significant. These results suggest that although conservative bloggers tend to more actively comment on one another's posts, this behavior is not accompanied by a greater uniformity in other online content they link to.

Besides looking at the citations bloggers make, we can also compare the similarity in the textual content of their blogs. Conservative television programs and conservative talk radio have sometimes been perceived to be acting as an echo chamber for Republican talking points. However, we did not find evidence for this in conservative blogs. To compare posts textually, we extracted a set of informative phrases, for example, "forged documents" or "vice presidential debate."

The set of informative phrases was extracted using a phrase finding algorithm which identifies phrases that are most informative with respect to a background model of term frequencies in weblog data. The first step in the algorithm identifies key *bigrams* in our corpus of weblog terms. The algorithm for finding key bigrams combines a measure of *informativeness* and a measure of *phraseness* for a bigram into a single unified score to produce a ranked list of key bigrams [16]. Next, the phrase finding algorithm finds all frequent phrases that contain any of the top N ranked bigrams and satisfy a set of phrase boundary tests.

We identified 498 such phrases across the 40 blogs, with each blog typically using a few hundred of the phrases. We then computed a cosine similarity measure between all pairs of blogs, this time using a $TF * IDF$ metric, where the entry in x_A corresponding to phrase p is given by $f_{A,p} * log(N/n_p)$, where $f_{A,p}$ is the number of times the phrase p occurs in blog A, $N = 1,768,887$ is the number of blogs harvested by BlogPulse between Oct.-Nov. 2004 and n_p is the number of blogs mentioning phrase p in all of the BlogPulse dataset. Interestingly, we found that it was the liberals who had a slightly higher pairwise similarity in the phrases they mentioned. As one would expect, the average similarity between blogs of opposite persuasions was smaller (0.10) than that of liberal (0.57) and conservative (0.54) pairs. So at first glance, we do not see evidence of a Republican "noise machine" at work in the blogosphere.

3.3 Interaction with mainstream media

Even more common than links to other blogs are links to news articles. Overall, the 20 left leaning bloggers cited the media 6,762 times, while the top 20 right leaning bloggers cited media 6,364, or, on average, about once every other post.

Figure 4 shows the most popular online news sites, and the proportion of liberal and conservative blogs linking to them within the top 20 liberal and the top 20 conservative blogs. As our analysis of the home pages of the larger set of political blogs will show in Section 3.5, we find that Fox News and the National Review receive the majority of their links from the conservative weblogs, while Salon receives over 86% of its links from liberal blogs.

Within the set of top political blogs, we also find that the NY Post, the WSJ Opinion Journal and the Washington Times receive the large majority of their links from right leaning blogs, while the LA Times, the New Republic and the Wall Street Journal are predominantly linked to by left leaning blogs. The remaining top-linked media sources are fairly evenly cited by the left and the right.

The actual news article citation behavior of the A-List political bloggers further differentiates the media sources attended to by bloggers on opposite sides of the political spectrum. Drilling down, here are the top news articles cited by left leaning bloggers:

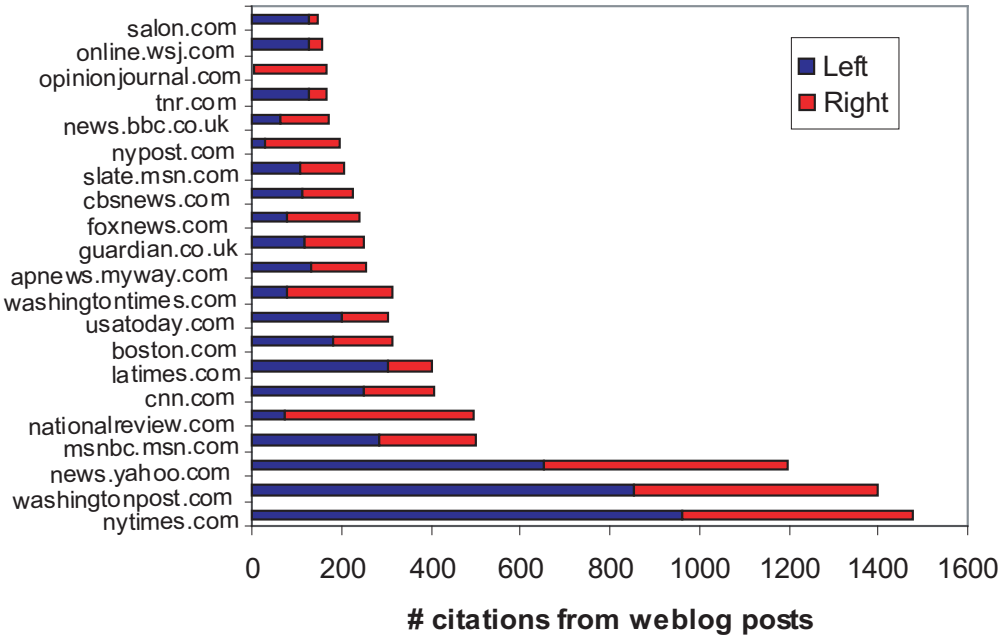

Figure 4: Most linked to news sources by the top 20 conservative and top 20 liberal blogs during 8/29/2004 - 11/15/2004.

1. CBS News poll of uncommitted voters shows Kerry winning 43% to 28%
2. Sun Times article: Bob Novak predicts that George Bush will retreat from Iraq if reelected
3. CBS News article on forged memos
4. New York Daily News article on Osama Bin Laden videotope, "gift" for the President
5. Time Magazine poll: Bush opens double-digit lead on post convention bounce

In contrast, the top news articles cited by right leaning bloggers are:

1. CBS News article on forged memos
2. Time Magazine poll: Bush opens double-digit lead on post convention bounce
3. National Review article refuting the case about missing explosives
4. ABC News article refuting the case about missing explosives
5. Washington Post article reporting on Kerry's proposal to allow Iran to keep its nuclear power plants in exchange for giving up the right to retain the nuclear fuel that could be used for bomb-making

A time series chart further shows how quickly and strongly conservative bloggers responded to forged CBS documents (Figure 5). The conservative bloggers saw Dan Rather's report as an attempt by the left to discredit President Bush. They acted quickly to debunk the report, with the charge led by PowerLine and seconded by Wizbangblog and others. In contrast, the pick-up among liberal bloggers occurred later, with lower volume. The most vocal left leaning bloggers on the subject were TalkLeft and AMERICAblog.

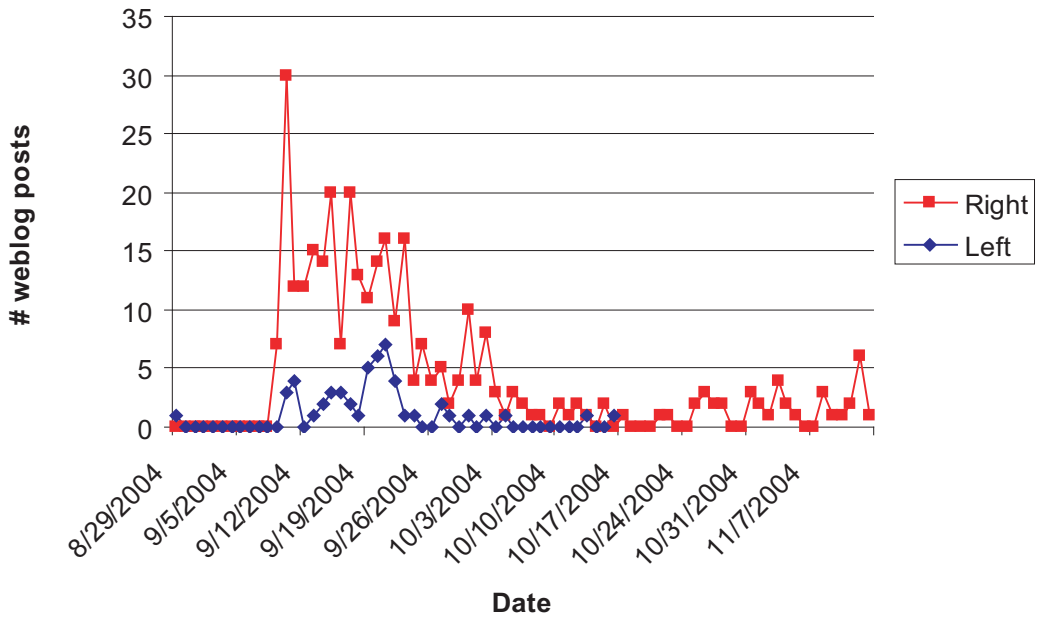

Figure 5: Time series chart: # posts discussing the CBS forged documents, right vs. left.

3.4 Occurrences of names of political figures

Figure 6 shows the relative number of mentions of the most cited political figures in our set of top 40 political weblogs. This chart excludes George W. Bush and John Kerry, with 8071 and 8011 mentions, respectively, in order to improve readability. Interestingly enough, the right leaning bloggers account for 59% of mentions of John Kerry, while the left account for 53% of mentions of George Bush.

The chart shows that some political figures are the focus of attention of primarily one side of the political spectrum. For example, the following figures are cited by name predominantly by the right: Dan Rather, Michael Moore, Yassar Arafat and Terry Mcauliffe. On the other hand, the left leaning bloggers account for most mentions of: Donald Rumsefeld, Colin Powell, Zell Miller and Tim Russert.

Notice the overall pattern: Democrats are the ones more often cited by right-leaning bloggers, while Republicans are more often mentioned by left-leaning bloggers. (While Zell Miller is officially a Democrat, he spoke at the Republican Convention and has been outspokenly anti-Kerry). These statistics indicate that our A-list political bloggers, like mainstream journalists (and like most of us) support their positions by criticizing those of the political figures they dislike. An interesting topic for further study would be to compare how balanced bloggers' presentation of the facts are compared with that of mainstream media journalists.

To create this data, we ran a person name extractor over the sets of left-leaning weblog posts and right-leaning weblog posts and counted up occurrences of the same name. We excluded the names of weblog authors for our top 40 weblogs. We then grouped together by hand different variants of the same name. The person name extractor has a recall of about 90%, so it is to be expected that one to three major political figures are missing from the ranking.

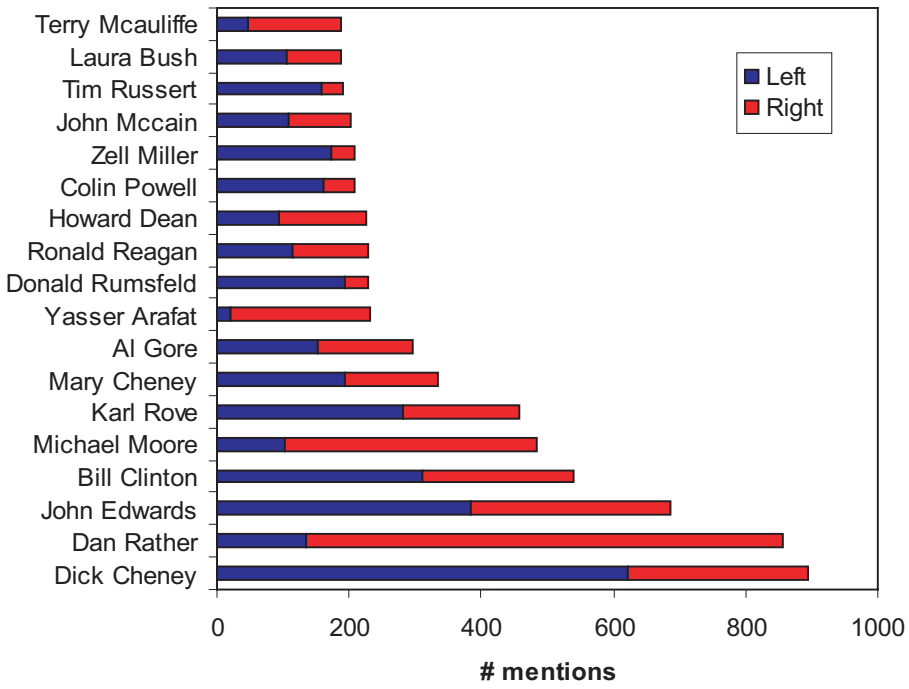

Figure 6: Mentions of political figures in liberal vs. conservative weblogs (excludes George W. Bush and John Kerry)

3.5 Back to the Greater Political Blogosphere

In addition to the A-list political blogs, there are hundreds of other blogs in the political blogosphere. We compared their interests in mainstream media based on the links extracted from their pages, and found these closely mirrored those of the top 40 blogs. Figure 7 shows just the most popular online news sites, and the proportion of liberal and conservative blogs linking to them. We see the same biases in the preferences of the large network as we did for the sample of most popular blogs. Fox news and the National Review receive 89% and 92% of their links, respectively, from conservatives, while Salon receives 91% of its links from liberal blogs. The New York Times, Google News, and the Washington Post are the most balanced, with the Washington Post being slightly favored by conservatives (55 to 40), while the NYT was almost even, with 5 more liberals than conservatives linking to it. One difference to note is that the NYT, Washington Post and Yahoo News receive a proportionally much greater share of the citations from posts than from blogroll links. For example, Salon receives about 60% as many blogroll links as does the New York Times. But it only receives 11% as many citations from blog posts. This suggests that some news sources often linked to by bloggers on their sidebars are read less frequently or spur less discussion.

Links to non-blog and non-news sites reveal other preferences that the two categories of blogs have. In what follows, there are two numbers next to each site, the first being the number of inlinks from liberal blogs, the second being the number of inlinks from conservatives. Conservatives get their daily cartoon fix at conservative cartoonist blogs 'Day by Day'(1,48) and 'Cox and Forkum'(1,83), while liberals get their laughs at the Onion (50,14). Liberals organized around thereisnocrisis.com (85,2) to defend social security against the plans of President Bush. They also link to MoveOn.org (56,9), and lend an ear to Michael Moore (44,10) and Tom Paine (41,1). The conservatives pay attention to the Middle East Media Research Institute (3,37), and listen to Sean Hannity (2,39)

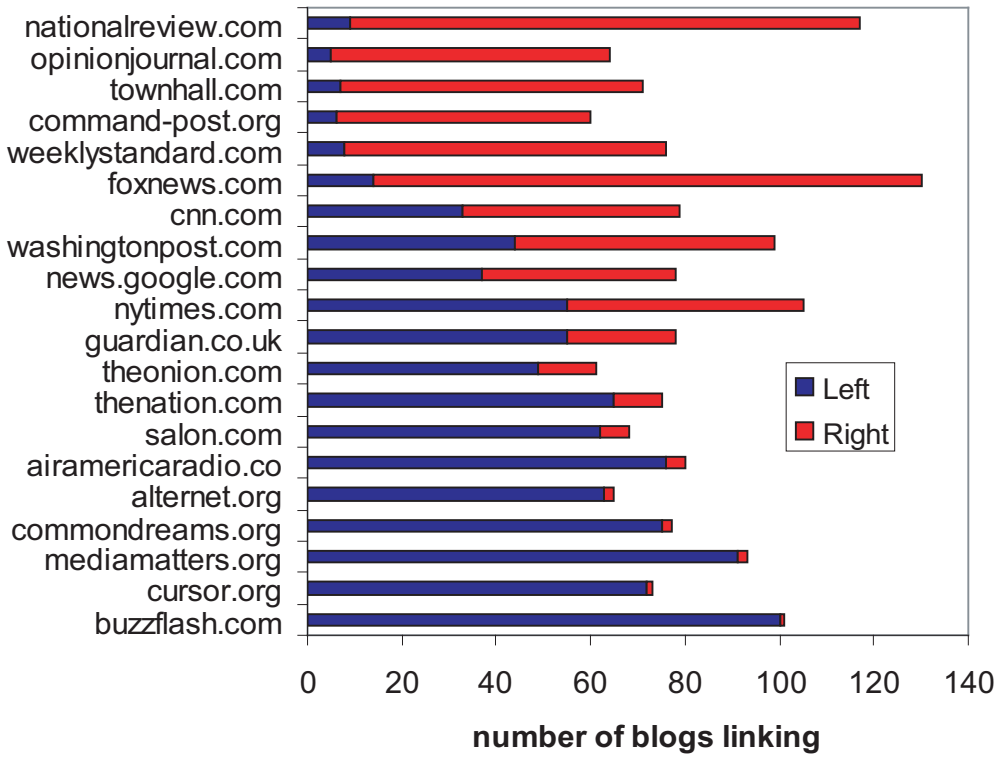

Figure 7: Most linked to news sources (online and off), showing proportionally how many liberal and conservative blogs link to them.

and Rush Limbuagh (3,29). They link to the GOP site (0,37) and are influenced by conservative think-tanks: the Heritage Foundation (1,33) and the Cato Institute (3,25).

4 Conclusions and Future Work

In our study we witnessed a divided blogosphere: liberals and conservatives linking primarily within their separate communities, with far fewer cross-links exchanged between them. This division extended into their discussions, with liberal and conservative blogs focusing on different news articles, topics, and political figures. An interesting pattern that emerged was that conservative bloggers were more likely to link to other blogs: primarily other conservative blogs, but also some liberal ones. But while the conservative blogosphere was more densely linked, we did not detect a greater uniformity in the news and topics discussed by conservatives.

There are still many questions that we would like to explore regarding the political blogging community. Some would involve extending our current methods, and others would take us further into the intricate behavior of the blogosphere. As a refinement of our current approach, we would like to differentiate between blogs written by a single author and those written by multiple contributors. Our experiments treated individual blogs as if they were each written by one author with a single voice, point of view and patterns of language use and linking behavior. However, while single author weblogs are more common than collaborative blogs, a number of the blogs in our two top 20 lists have (or had) multiple authors, for example: blog.johnkerry.com, crookedtimber.org, pandagon.net, deanesmay.com, powerlineblog.com, blogsforbush.com, dailykos.com, wiz-

bangblog.com, and www.prospect.org/weblog. In future work, we would like to repeat our cosine similarity measurements for links and phrases over unique authors instead of over individual weblogs. Even more ambitious, from a data collection and information extraction point of view, would be to identify the most cited weblog authors and redo our experiments over a collection of posts segmented by author instead of by weblog URL.

We would also like to track the spread of news and ideas through the communities, and identify whether the linking patterns in the network affect the speed and range of the spread. For example, we observed that the CBS document discussion was much more active in the conservative blogosphere. Was it because of stronger interaction patterns of the conservatives, or did the liberals simply not want to discuss it?

Finally, there is the matter of the 'other' political blogs, those calling themselves 'independent' or 'moderate'. Very few popped up on our radar, with none making our top 20 lists. At the very least, however, we could see whether they act as bridges between the liberal and conservative communities, or if they form their own community which may be just as isolated. Will they grow in number and start gaining in popularity as the divisive 2004 U.S. Presidential election fades from memory? Any change in the balance and interaction of the political blogosphere makes for an interesting subject of study.

5 Acknowledgments

We would like to thank Eytan Adar for insightful discussions and TJ Giuli and Marita Silverstein for helpful comments and suggestions. We would also like to thank the BlogPulse team, especially Matt Hurst and Mark Reed for their vital technical contributions, Sundar Kadayam for his vision and guidance, and Sue MacDonald for her inspired editorial analyses of the political blogosphere.

References

[1] L. Adamic and B. Huberman. The nature of markets in the world wide web. *QJEC*, 1:5–12, 2000.

[2] L. A. Adamic. The small world web. In *Proceedings of the 3rd European Conf. on Digital Libraries*, volume 1696 of *Lecture notes in Computer Science*, pages 443–452. Springer, 1999.

[3] E. Adar. Guess: The graph exploration system. `http://www.hpl.hp.com/research/idl/projects/guess/guess.html`, 2005.

[4] P. Boutin. Net-savvy campaign boosts bush. In *WIRED*. `http://www.wired.com/news/politics/0,1283,63942,00.html`, 2004.

[5] D. W. Drezner and H. Farrell. The power and politics of blogs. `http://www.danieldrezner.com/research/blogpaperfinal.pdf`, 2004.

[6] N. Glance, M. Hurst, and T. Tomokiyo. BlogPulse: Automated trend discovery for weblogs. In *WWW 2004 Workshop on the Weblogging Ecosystem: Aggregation, Analysis and Dynamics*, 2004.

[7] S. C. Herring, I. Kouper, J. C. Paolillo, and L. A. Scheidt. Conversations in the blogosphere: An analysis "from the bottom up". In *HICSS-38*. Springer, 2005.

[8] M. Hindman, K. Tsioutsiouliklis, and J. A. Johnson. "googlearchy": How a few heavily-linked sites dominate politics on the web. `www.princeton.edu/~mhindman/googlearchy--hindman.pdf`, 2004.

[9] V. Krebs. The social life of books, visualizing communities of interest via purchase patterns on the www. `http://www.orgnet.com/booknet.html`, 2004.

[10] C. Marlow. Audience, structure and authority in the weblog community. In *International Communication Association Conference*, New Orleans, LA, 2004. `http://web.media.mit.edu/~cameron/cv/pubs/04-01.pdf`.

[11] J. Newsome and D. Song. Gem: Graph embedding for routing and data-cenric storage in sensor networks without geographic information. In *SenSys'03*. ACM, 2003.

[12] D. M. Pennock, G. W. Flake, S. Lawrence, E. J. Glover, , and C. L. Giles. Winners don't take all: Characterizing the competition for links on the web. *Proceedings of the National Academy of Sciences (PNAS)*, 99(8):5207–5211, 2002.

[13] R. Putnam. *Bowling Alone*. Simon and Schuster, New York, 2000.

[14] C. Shirky. Power laws, weblogs and inequality. `http://shirky.com/writings/powerlaw_weblog.html`, 2003.

[15] C. Sunstein. *republic.com*. Princeton University Press, Princeton, 2001.

[16] T. Tomokiyo and M. Hurst. A language model approach to keyphrase extraction. In *Proceedings of the ACL Workshop on Multiword Expressions*, 2003.

[17] P. Welsch. Revolutionary vanguard or echo chamber? political blogs and the mainstream media. Sunbelt 2005 presentation `http://www.blogninja.com/sunbelt05.pete.ppt/`, 2005.