2011

An automated snowball census of the political web

Abe Gong

University of Michigan, agong@umich.edu

Follow this and additional works at: <https://scholarworks.umass.edu/jitpc2011>

Part of the [Communication Commons](#), [Computational Engineering Commons](#), [Computer Engineering Commons](#), [Political Science Commons](#), [Public Affairs, Public Policy and Public Administration Commons](#), and the [Science and Technology Studies Commons](#)

Gong, Abe, "An automated snowball census of the political web" (2011). *JITP 2011: The Future of Computational Social Science*. 6.
Retrieved from <https://scholarworks.umass.edu/jitpc2011/6>

This Article is brought to you for free and open access by the The Journal of Information Technology and Politics Annual Conference at ScholarWorks@UMass Amherst. It has been accepted for inclusion in JITP 2011: The Future of Computational Social Science by an authorized administrator of ScholarWorks@UMass Amherst. For more information, please contact scholarworks@library.umass.edu.

An automated snowball census of the political web

Abe Gong
Working paper*

May 10, 2011

Abstract

This paper solves a persistent methodological problem for social scientists studying the political web: representative sampling. Virtually all existing studies of the political web are based on incomplete samples, and therefore lack generalizability. In this paper, I combine methods from computer science and sampling theory to conduct an automated snowball census of the political web and constructs an all-but-complete index of English political websites. I check the robustness of this index, use it to generate descriptive statistics for the entire political web, and demonstrate that studies based on ad hoc sampling strategies are likely to be biased in important ways. In future research, this bias can be eliminated by using this index as a sampling universe. In addition, the methods and open-source software presented here can be used to creating similar sampling frames for other online content domains.

Keywords: Sampling theory, web mining, text classification, computational social science

*Special thanks to Walter Mebane, Rob Franzese, Ted Brader, Paul Poast, and Andrea Jones-Rooy for comments on early drafts of this paper. Thanks also to Mike Bommarito, Matt Simmons, and Lada Adamic for technical guidance on python development and web crawling.

1 Introduction

I begin with a pressing problem, captured in quotations. “In the absence of a known population, ... a truly random sample [of relevant web sites] is not possible.” (Miller, Pole and Bateman, 2011) “Obtaining a random sample of bloggers is imperfect because of the amorphous nature of the blogosphere.” (McKenna and Pole, 2008) “The vast amount of human knowledge encoded online is the reason why the Web is such a valuable resource for politics; but ironically, the very scale of this resource makes the Web extraordinarily difficult to study.” (Hindman, Tsioutsoulis and Johnson, 2003).

Behind all of these quotations is an unstated understanding: without a complete index of political web sites, social scientists studying the web are denied one of their most powerful tools—sampling theory. Sampling theory is the keystone of an enormous body of social science research. It enables us to draw conclusions from manageable samples, and generalize them to whole populations.

On the web—where we have lacked a valid sampling frame—we have been unable to make such claims about representative sampling and generalizability. Consequently, many otherwise compelling studies of online behavior have a piecemeal flavor. (See section two for examples.) Although they may discover interesting and important patterns among the bloggers, media sites, forum participants, online advertisements, etc. chosen for study, they cannot say for certain whether those patterns are representative of the political web at large.

In this paper, I show how to solve the web-sampling problem for a broad class of applications. Using a novel recombination of methods from computer science and sampling theory, I construct a comprehensive index of English political websites. I check the robustness of this index, use it to generate descriptive statistics for the entire political web, and demonstrate that studies based on ad hoc sampling strategies are likely to be biased in important ways. I hope that this index of nearly 800,000 web sites will facilitate future research using surveys and content analysis to understand the Internet and its politically minded inhabitants. Furthermore, I hope that the methods and software presented here will be useful for exploring and sampling from other domains on the web.

The paper proceeds as follows. Section two reviews past attempts to sample the political web, focusing on methodological limitations. Section three describes my methodology, in detail. Section four reports results, in-

cluding web crawling statistics and robustness checks. Section five discusses the strengths and limitations of this approach. Section six summarizes and concludes.

2 Literature Review

To the best of my knowledge, no previous study of political websites has been based on a fully representative sample. All past studies use convenience, prominence, snowball, or over- samples. Here I describe these common sampling strategies, give examples, and highlight their limitations.

2.1 Convenience sampling

In a convenience sample, no attempt is made to make the sample population representative of the population as a whole. Instead, researchers “look under the light post” by gathering a sample that is close at hand. For instance, several studies have used top results from search engines such as Google as the basis for their analysis. Others have relied on opt-in recruitment through pop-up and banner ads. Although studies of this kind may achieve high internal validity, they cannot draw generalizations about the population as a whole.

Examples of studies employing convenience samples include: Davis’ study of usenet discussion forums (Davis, 2009); Baum and Groeling’s analysis of news judgements on left- and right-leaning online news sites (Baum and Groeling, 2008); and Johnson and Kaye’s study of news readership among blog readers (Johnson and Kaye, 2004).

2.2 Prominence sampling

Prominence samples focus on the most visible sites according to some well-defined metric. Prominence samples can be used to draw conclusions about popular sites—the ones most likely to show up at the top of these rankings—but they can’t be used to make inferences about the political web in general. For example, several studies of political blogging have based their samples on lists of the most popular political blogs, according to tracking companies such as Technorati or Truth Laid Bear.

This method is very popular. Examples of studies employing prominence samples include: Adamic and Glance’s network analysis of ideological clustering in popular political blogs (Adamic and Glance, 2005); Davis’ work on political blogging (Davis, 2009); and McKenna and Pole’s survey of “A-list” political bloggers (McKenna and Pole, 2008). Perhaps the most thorough example is Wallsten’s content analysis of blog posts, which is based on a random sample of 10,732 sites featured on popular blog directories (Wallsten, 2008).

2.3 Snowball sampling

Snowball samples start from a batch of known sites, then “snowball out” to gather other sites for the sample. Conceptually, this approach has some advantages over prominence sampling, because it potentially includes all sites, not just the most prominent. However, snowball samples are still not random according to the definition of sampling theory. Although some attempts have been made to remedy this problem (Salganik and Heckathorn, 2004) (Rivers, 2007), we still lack universally accepted statistical tools for drawing inferences from snowball samples.

A few studies have used this methodology. See for example, Hindman et al.’s early work on power laws in political web sites (Hindman, Tsioutsoulouklis and Johnson, 2003); Ackland’s attempt to map network neighborhoods of elite blogs within the U.S. political blogosphere (Ackland, 2005); Karpf’s index of “authority” among prominent blogs (Karpf, 2008); and Miller and Pole’s recent content analysis of health blogs (Miller, Pole and Bateman, 2011).

2.4 Oversampling

Oversampling starts with a very large random sample from the general population. After a short screening interview, respondents matching certain criteria are selected into the final sample. Oversampling follows a straightforward statistical methodology, but the cost of gathering a large enough starting sample is often prohibitive. In addition, any bias in answers on screener questions (e.g. from recall, social desirability, or interview fatigue) can skew the composition of the resulting sample. Also, validating survey answers against actual online behavior is very difficult—to the best of my knowledge, no study to date has attempted to do so.

Lenhart and Fox (2006) provides a good example of this approach. Over the course of a year, they added a screener question about blogging to the end of 13,000 Pew phone interviews. The 233 respondents who said they blogged were later called back and interviewed at length about their blogging habits. This approach worked reasonably well for gathering bloggers in general, but to identify a statistically significant sample of political bloggers would have required a much larger sample—almost certainly too large for most research budgets.

Similar studies include Schlozman et. al’s study of online participatory inequality (Schlozman, Verba and Brady, 2010), and Lawrence and Sides’ survey analysis of blog readership (Lawrence, Sides and Farrell, 2010).

2.5 Summary

In summary, past studies of the political web have not been based on representative sampling methods. Although many of these studies achieve high internal validity, the lack of a universal sampling frame makes it extremely difficult to achieve high external validity. Consequently, no existing study has been able to draw generalizable conclusions about the composition of the political web as a whole.

3 Methodology

My goal in this paper is to solve that problem, subject to the statistical requirements of sampling theory and the technical constraints imposed by the enormous scale of the web.

The intuition behind my approach is straightforward: all publicly available websites can be accessed over the Internet, and virtually all websites are connected by links among webpages. In principle, we should be able to build a sampling frame for the political web by following links among web sites, and classifying them one by one.

In principle, this approach works. The practical problem is time. Experts estimate that 255 million web sites existed as of 2010 (royal.pingdom.com, 2011). To borrow a phrase from Matthew Hindman, exploring so many sites would be the work “of many lifetimes.” To illustrate, roughly 59,000 new websites are started every day. Some percentage of those are political. Just keeping pace with these new additions would require reading 40 new sites

every minute, forever. With so many sites to search through, constructing a sampling frame by hand is not feasible.

3.1 An automated snowball census

Fortunately, this process can be automated. Instead of searching through millions of websites in person, we can write software to search through millions of websites for us. The software combines two common tools from computer science: web spiders and text classifiers.

A *web spider* is a program that explores and downloads online content by following links on web pages. Thus, the spider simulates the process of surfing the Internet by clicking one link after another. Spiders can do this tirelessly, and—when properly designed—very fast. With a good Internet connection and parallel threaded architecture, a spider can easily “crawl” dozens or hundreds of web pages in a second.

A *text classifier* is a text-as-data algorithm that categorizes documents based on the words that appear in them. Text classifiers have been used for information retrieval and natural language processing (NLP) for many years (Maron, 1961), but their use has dramatically accelerated with the rise of the Internet (Manning et al., 2008). As I will explain shortly, text classifiers can be adopted to the needs of social scientists, yielding high-reliability content coding on a virtually unlimited scale.

Combining web spiders and text classifiers allows us to conduct an “automated snowball census” of the political web, as follows:

1. Start with a seed batch of likely political sites.
2. Download this batch, and classify each site as political, or not.
3. For each political site, harvest all the outbound hyperlinks.
4. Place every previously unvisited hyperlink in the next batch of sites to be visited.
5. Repeat from step two until no political new sites are found.

This approach follows the logic of snowball sampling: the best place to look for political sites is close to other political sites in the network. Unlike snowball *sampling*, this *census* follows that logic to the bitter end. Every

single site linked from at least one known political site is checked for political content. By cataloging the sites visited in the snowball search, we can create an index of political websites. My claim is that virtually every political website will be included in an index created this way.

This is a strong claim, resting on two key provisions. First, the classifier must classify accurately. Any false positives or false negatives could distort the sample. Second, political websites must be adequately connected. If the political web is fragmented into disjoint islands of content not connected by any links, the snowball might blanket one island but never reach the others.

Fortunately, good evidence supports my claim. The process of training the text classifier and spidering the political web reveals a great deal about the reliability of the classifier and connectedness of the network along the way. I describe these results in the next two sections.

3.2 Training a text classifier

My approach to training a text classifier combines best practice from traditional content analysis with recent innovations in natural language processing. The main idea is to define political content so that it can be reliably categorized by human readers, then use statistical techniques to train an algorithm to mimic human coding. I describe these steps here in turn.

I define political content as content “focus[ing] on who controls power in government, and/or how that power is used.” This definition works well in practice, lining up nicely with common intuition about political content, while also aligning with important theoretical concepts about the authority of the state (Burns, Schlozman and Verba, 2001). This definition excludes non-state power relationships, such as gender dynamics in the workforce.

Using the crowd-sourcing service Amazon Mechanical Turk, I recruited English-speaking U.S. residents to code 2,200 potentially political websites. For each site, coders were given instructions, a link to the website, and a simple form to fill out. Appendix A contains these instructions and codesheet. Coders were paid three cents per site, regardless of whether it was political or not. In addition, 200 sites were coded four times each, in order to check inter-coder reliability. When coders disagreed, I used the median value as the final code for the site. These 2,200 labeled sites served as training and testing data for the classifier.

For the classification algorithm itself, I used regularized logistic regression (RLR) over a set of 2,000 maximally informative word stems. Essentially, this

classifier counts the occurrences of common words with strong associations (or disassociations) to politics, and uses a kind of weighted average to decide whether the site is political or not. Based on the training data, the classifier assigns a score to each word. For example, words like “Obama,” “Senate,” and “vote” are strongly associated with political content, and therefore receive large positive scores. The words “film,” “home,” and “museum” are weakly disassociated with political content, and therefore receive moderate negative scores¹ By aggregating scores over all the words in the document, the classifier can arrive at a very good guess as to whether the document as a whole is political. Appendix B illustrates the final political classifier with a colored word cloud.

Using RLR for classification has several advantages. First, RLR has been shown to perform as well or better than other state-of-the-art class text classification algorithms on a variety of classification tasks. Second, fast algorithms for training regularized logit classifiers have been implemented in publicly available software packages. I used the python library `scikits.learn`, which wraps the `LIBLINEAR` library for fast classifier optimization. Finally, RLR has a strong statistical foundation. With a small regularization constant, RLR asymptotically approaches the maximum likelihood linear hyperplane classifier as sample size increases. See Zhang and Oles (2001) for a detailed technical exposition of these issues²

These theoretical justifications are reassuring, but the best proof of classifier effectiveness must come from tests on real data. Regardless of theoretical credentials, a classifier that makes too many mistakes cannot be trusted to categorize the whole political web. No classifier is perfect, and mine is no exception. When evaluated against human coders, my classifier for political content agrees 80.1% of the time, coding 34% of documents as political³. A naive reading of these results is that the classifier is “about 80% accurate.” However, this puts the classifier in an unfairly negative light. When com-

¹Note that the strength of association for each stem is estimated conditional on all others. This is not a “naive” classifier.

²As discussed by Zhang and Oles, one disadvantage of RLR is that the researcher must choose the regularization constant without much theoretical guidance. I experimented with constants on the range $[10^{-10}, 100]$ and finally chose 10^{-5} because it seemed to give a small boost to classifier accuracy.

³Following best practice in NLP, all accuracy and reliability statistics were generated by applying the classifier to a testing set of documents separate from the training set. This approach reduces the risk of overfitting.

pared to each other, the same human coders agree only 80% of the time⁴. The computer agrees with the human coders more than the human coders agree with each other!

The reason lies in the definition of “political” content. When applied to the messiness of the real world, even the most clear and crisp definition has some gray area where reasonable coders can disagree. This gray area accounts for most of the difference between human coders. The computer does slightly better than the human coders because the process of training leads it to balance across the coding styles of many humans. So not only is automated coding dramatically faster than human coding, it’s also a shade more accurate as well⁵.

Training data and source code for this classifier are available on request.

3.3 Homophily

The second potential problem is isolated sites. If a given site or set of sites was disconnected from the rest of the political web, the snowball spider would never find it. My first defense against this claim comes from network theory. Like many social networks, the political web exhibits a high degree of homophily. Political sites are much more likely to link to political sites. More precisely, roughly 1 in 3 links from political sites are directed to other political sites, despite the fact that only about 1 in 300 websites is political. Links among websites are not random; the best place to look for links to political sites is from political sites.

A second defense comes from the scope of the research project: for all intents and purposes, isolated sites are not part of the public sphere. No other political site links to them. If they have no in-links, then no search engine can index them. Without search engine traffic or links, the only people who could visit such sites are those who already know their exact web addresses. Therefore, it seems fair to say that posting to an isolated site is a private action, not an act of public political participation.

⁴These results are from reliability tests performed on a random sample of sites encountered in the snowball census. Therefore, they best represent sites in or near the political region of the World Wide Web. For the web as a whole, the reliability of both humans and computers would probably improve.

⁵The difference is easily within the margin of error, so it may be more fair to say that the computer codes *just as accurately* as human coders.

4 Summary

So far, I have described a process for conducting an automated snowball census of the political web. Conceptually, the same approach should work for any subdomain on the web, as long as (1) the text classifier is accurate, and (2) the network is sufficiently connected. The political web seems to meet both of these criteria. In the next section, I describe results from the process and supply additional robustness checks.

5 Results

I implemented the process described above in python. SnowCrawl, an open-source python module, provides a common API for directed webcrawls using a single process, multiple processing, or a client-server architecture. SnowCrawl also automates storage of downloaded files, edge lists, and state backup. Source code, examples, and documentation are available on google code <http://code.google.com/p/snowcrawl/>.

For a snowball census conducted in multiprocessing mode August 2010, the code executed in less than 24 hours, crawling some 1.8 million sites, and classifying about 800,000 as political.

5.1 Robustness checks

Does it actually work? Without another census to compare against, comprehensive tests are impossible. However, we can run some “back-of-envelope” validity checks.

First, we can ask if the web spider found about the right number of political sites. Older work (Hindman, 2010) based on patterns of browsing traffic on the Internet, placed the percentage of political sites around a third of a percent. Given last year’s estimate (royal.pingdom.com, 2011) of 255 million web sites and monthly growth of 7.1 million sites, we should expect to find about 826,000 political sites in our August crawl. The total count from my snowball census is in just the right ballpark: 789,818 political sites.

As a second robustness check, we can look to see if any known political sites are obviously missing or misclassified. Prominent political sites show up early in the sample: The Huffington Post, Daily Kos, whitehouse.gov, and so on. Appendix C lists the top 200 political web sites, in no particular

order. A quick search within the index also reveals 497 house.gov sites and 217 senate.gov sites in the census. These are the official websites of U.S. Congressmen, Senators, and committees—it appears that the classifier found all of them.

A third robustness check considers the network characteristics of the political web. I have already discussed homophily: proportionally speaking, political sites are much more likely to link to other political sites. It follows that links between political sites are much denser than links between political sites and the rest of the web.

Taken together, these checks provide reasonably strong evidence that this index is close to complete. As a sampling universe, it is almost certainly superior to previous studies of the political web.

5.2 Descriptive statistics

With this index in hand, we can generate the first fully representative description of political web. To do so, I sampled 150 websites from each of three strata within the census: the top 500, top 5,000, and full census of political websites. Strata were determined by inlinks from sites found in the crawl⁶. For each site, workers on Mechanical Turk were paid \$0.10 to fill out a short codesheet (see Appendix D for details) including questions about the ownership, organization, and content of the site.

Table 1 reports descriptive statistics for each stratum. For the purposes of describing the the political web, the important data are in the rightmost column, which reports percentages for a sample of the full census. Thus we see that 55.6 percent of political websites are personal websites run by individuals or informal groups, as opposed to websites run by organizations such as news media outlets, political campaigns, corporations, etc. 59.5 percent of political sites are formatted as blogs; 62.2 percent have more than one author; only 6.1 percent are updated multiple times per day.

I also asked about certain design elements within pages. Half of sampled sites included advertising, and nearly half included a “blogroll” or collection of links to related sites. About one in five political sites features video content. Forty percent include identifying information about their authors. Somewhat surprisingly, nearly a third of site include buttons or forms solic-

⁶As Hindman (2010) has shown, inlinks are strongly correlated with other measures of popularity, such as traffic and search engine rankings.

Table 1: Characteristics of political websites by strata

	Top 500	Top 5,000	Census
<i>Organization</i>			
Personal websites	33.9%***	46.9%	55.6%
Owned by organizations	66.1***	53.1	44.4
Formatted as blogs	51.4	70.5	59.5
Multiple authors	75.2*	66.7	62.2
Multiple updates per day	43.4***	19.4***	6.1
Updated less than weekly	14.2***	21.4***	42.7
<i>Design</i>			
Advertising	67.3**	57.1	51.2
Blogroll	57.5*	66.3***	45.1
Videos	48.7***	35.7***	18.3
Identifying information	47.8	50.0	41.5
Forms for donations, etc.	36.3	32.7	30.5
<i>Content</i>			
Polls and public opinion	70.8***	65.3*	52.4
Elections and campaigns	50.4	45.9	51.2
Legislation and law-making	43.4	41.8	43.9
Implementation of policy	38.1	39.8	30.5
Decisions by courts	34.5***	24.5	17.1
Political figures	46.0***	39.8**	24.4
Political parties	38.9***	32.7*	20.7
Philosophical discussion	26.5	29.6	25.6
State and local government	36.3*	38.8**	24.4
Foreign policy	42.5***	38.8***	15.9
International relations	31.9**	33.7**	18.3

Cell entries show the percent of sites that have certain organizational characteristics, or contain design element or content types. Stars indicate significance levels in pairwise t-tests between Census results and the Top 500 or Top 5,000 strata (* $p < .1$)(** $p < .05$)(*** $p < .01$). Many aspects of political websites differ significantly between sites in the head and tail of the distribution.

iting donations or recruiting volunteers.

In terms of political content, polls, public opinion, and elections appear to be the most popular topics, followed by legislation and law-making, implementation and execution of public policy, philosophical discussion of the role of government, state and local government, political figures, and political parties. In general, foreign policy, international relations, and decisions by courts received less attention on political sites.

5.3 Comparison with previous methods

Since all previous studies have used less than fully representative samples, it is reasonable to ask what difference, if any, a more complete sampling frame would have made. The other columns in Table 1 allow us to make comparisons between popular sites in the head of the distribution, the entire population of English-language political websites. Values that differ significantly from the census value (as measured by pairwise t-tests) are marked with asterices.

Thus, we see that the top 500 websites are more likely to be owned by organizations and maintained by multiple authors, with far more frequent content updates. Popular sites are also more likely to feature ads, links to other relevant sites, and video content. In this sample, they were slightly more likely to solicit donations and volunteers, but the difference between strata was not statistically significant.

A-list sites also differ in the kinds of content they cover. Popular sites include more types of content overall: 4.74 types for the top 500, and 4.50 for the top 5,000, versus 3.46 for the full census. Most of the difference comes from increased coverage of foreign policy, political figures, polls and public opinion, and court decisions. These differences in attention are substantively large: compared to sites in the full census, a top 500 site is twice as likely to discuss decisions by courts and nearly three times as likely to discuss foreign policy.

These results underscore the importance of proper sampling. Without a proper sampling frame, substantive conclusions about the political web are likely to be biased—in some cases, severely.

6 Discussion

This section discusses contributions, limitations, and directions for future research.

6.1 Contribution

The index created here should enable representative sampling of the political web. Generating the index takes specialized software and computing power, but once generated, it can be used as needed, with little technical expertise. Like a phonebook provides an off-the-shelf method for sampling in phone surveys, this index provides an easy way to sample from the political web. The full census and various sub-samples of interest are available for download at <http://www-personal.umich.edu/~agong/resources.html>.

Furthermore, researchers with the requisite programming expertise have the option of running the snowball software again. In this case, different classifiers can be trained and used to explore different subdomains on the World Wide Web.

6.2 Limitations

Like all methods, the automated snowball census described here has some potential limitations. In particular, mistakes in coding and isolated sites in the network might threaten the validity of the census.

As described above, the classifier described here performs quite well—better than human coders. However, there are several means by which incremental improvements in classifier accuracy might be obtained. First, better instructions and training for human coders could eliminate some errors in the training data used to calibrate the classifier. This approach would likely improve both human-human and human-computer reliability scores. Second, in a similar vein, additional training data would likely lead to modest improvements in classifier accuracy. Third, an expanded feature space including more words or perhaps bigrams and trigrams, would probably improve the classifier a little bit as well. Fourth, experimentation with different classifiers (e.g. nonlinear SVM kernels) might also improve the classifier slightly. Overall, given the already-high accuracy of the political classifier, we should expect incremental improvements in performance at best.

In a more promising direction, the classifier could be redesigned to incorporate other forms of information. At present, the classifier only makes use of text on the main page of a web site. Consequently, short pages (i.e. those containing less than 100 words) offer less material for classification and are more likely to be misclassified. With some additional effort, the classification algorithm could be trained to incorporate text from other pages within the site, the structure of the hyperlink network surrounding the site, and so on. For sites with few words, these additional information sources might substantially improve classification accuracy.

6.3 Directions for future research

The preliminary analysis offered in this paper also suggests several promising avenues for future research. First, rich description of the political web. With a sampling frame in hand, it should be easy to draw a sample, measure the properties of web pages and their authors using content analysis and/or surveys, then make generalizable inferences to the political web as a whole. This paper includes preliminary results in this direction, but much work remains to be done.

Second, investigation of organizational structure within sites. Results presented here hint at the diversity of organizational structures in the political web. Different sites have different ownership structures, design elements, and means of content production. Understanding the In addition to a wide array of design elements

Third, a more thorough study of network properties. The homophily and edge density statistics presented here offer only a cursory analysis of the structure of the political web. Because the snowball spider automatically generates an edge list for all sites searched, it opens exciting possibilities for studying the political web as a complex network. Such research could add greatly to our understanding of information flow, social connectivity, and political involvement.

Fourth, stability over time. The snowball census takes a snapshot of the political web. It would be interesting to see how the properties of that network and its constituent sites change from week to week.

7 Conclusion

In summary, this paper demonstrates how a combination of tools from computer and social science can be used to conduct an automated snowball census of the political web. I have argued that this process generates an all-but-complete index of the political web, providing strong theoretical and empirical support for that claim. Consequently, it seems reasonable to use the resulting index as a sampling universe for the political web, solving a persistent methodological problem in recent social science.

References

- Ackland, R. 2005. Mapping the US political blogosphere: are conservative bloggers more prominent? In *BlogTalk Downunder 2005 Conference, Sydney*. BlogTalk Downunder 2005 Conference, Sydney.
- Adamic, L.A. and N. Glance. 2005. The political blogosphere and the 2004 US election: divided they blog. In *Proceedings of the 3rd international workshop on Link discovery*. ACM pp. 36–43.
- Baum, M.A. and T. Groeling. 2008. “New media and the polarization of American political discourse.” *Political Communication* 25(4):345–365.
- Burns, N., K.L. Schlozman and S. Verba. 2001. *The private roots of public action: Gender, equality, and political participation*. Harvard Univ Pr.
- Davis, R. 2009. *Typing politics: the role of blogs in American politics*. Oxford University Press, USA.
- Hindman, M. 2010. *The myth of digital democracy*. Princeton University Press.
- Hindman, M., K. Tsioutsoulouklis and J.A. Johnson. 2003. Googlearchy: How a few heavily-linked sites dominate politics on the web. In *annual meeting of the Midwest Political Science Association*. Vol. 4 Citeseer pp. 1–33.
- Hopkins, D.J. and G. King. 2010. “A method of automated nonparametric content analysis for social science.” *American Journal of Political Science* 54(1):229–247.

- Johnson, T.J. and B.K. Kaye. 2004. "Wag the blog: How reliance on traditional media and the Internet influence credibility perceptions of weblogs among blog users." *Journalism and Mass Communication Quarterly* 81:622–642.
- Karpf, D. 2008. "Understanding blogspace." *Journal of Information Technology & Politics* 5(4):369.
- Lawrence, E., J. Sides and H. Farrell. 2010. "Self-segregation or deliberation? Blog readership, participation, and polarization in American politics." *Perspectives on Politics* 8(01):141–157.
- Lenhart, A. and S. Fox. 2006. "Pew Internet Report–Bloggers: A portrait of the internet\'s new storytellers."
- Manning, C.D., P. Raghavan, H. Schütze and Ebooks Corporation. 2008. *Introduction to information retrieval*. Vol. 1 Cambridge University Press Cambridge, UK.
- Maron, ME. 1961. "Automatic indexing: an experimental inquiry." *Journal of the ACM (JACM)* 8(3):404–417.
- McKenna, L. and A. Pole. 2008. "What do bloggers do: an average day on an average political blog." *Public Choice* 134(1):97–108.
- Miller, E.A., A. Pole and C. Bateman. 2011. "Variation in Health Blog Features and Elements by Gender, Occupation, and Perspective." *Journal of health communication* 99999(1):1–24.
- Rivers, D. 2007. Sampling for web surveys. In *Joint Statistical Meetings*.
- royal.pingdom.com. 2011. "Internet 2010 in numbers."
URL: <http://royal.pingdom.com/2011/01/12/internet-2010-in-numbers/>
- Salganik, M.J. and D.D. Heckathorn. 2004. "Sampling and Estimation in Hidden Populations Using Respondent-Driven Sampling." *Sociological methodology* 34(1):193–240.
- Schlozman, K.L., S. Verba and H.E. Brady. 2010. "Weapon of the strong? Participatory inequality and the Internet." *Perspectives on Politics* 8(02):487–509.

- Wallsten, K. 2008. “Political blogs: Transmission belts, soapboxes, mobilizers, or conversation starters?” *Journal of Information Technology & Politics* 4(3):19–40.
- Zhang, T. and F.J. Oles. 2001. “Text categorization based on regularized linear classification methods.” *Information Retrieval* 4(1):5–31.

Appendix A: A word cloud representing a text classifier for political content

Orange words are associated with political content, and blue words are disassociated. The size of a word denotes the strength of association -- essentially, the size of each word corresponds to the absolute value of the beta value of the word in a logistic regression with "political-ness" as the dependent variable. The layout of the words is done by computer algorithm to conserve space; it doesn't carry any important information.

Code sheet: Political content on web sites

Political content focuses on **who controls power in government, and/or how that power is used**. It includes content about elections, public opinion, public policy, public officials, and so on. State and local politics are included, along with foreign policy and relationships between countries. Personal ideas, opinions, and experiences can also be political, as long as the government is involved.

Warning: not all controversial issues are political. Whether or not we count them as political here depends on whether they **explicitly mention government involvement**. For example, discussions about the economy, environment, women's rights, etc. sometimes --- but not always --- match our definition of political content. When the issue is framed, is government part of the picture?

Not political

"The river by my house is polluted."

"There is still an income gap between men and women."

"Unemployment was up 2% last year."

"Terrorists bombed a night club in Bali."

"I don't believe in evolution."

Political

"I wish the city council would clean up the river by my house."

"There is still an income gap between men and women, but it's not the court's role to try and change that."

"Since the Federal Reserve increased the interest rate last year, unemployment has risen 2%."

"Terrorists opposing U.S. activity in the Middle East bombed a nightclub in Bali."

"I don't believe in evolution, so schools shouldn't teach it."

Please enter the site ID from the spreadsheet.

Please enter the site URL from the spreadsheet.

Please enter your unickname

Does the site load properly?

- ☐ Yes, the site loads immediately and I can read its contents.
- ☐ No, I get an error message like "Site not found," "Account has been suspended," etc.
- ☐ No, I get a site offering to redirect me to a different page.
- ☐ No, the site requires a password or invitation for access.
- ☐ No, the site fails to load for some other reason.

Is the site primarily in English?

- ☐ Yes, the site is entirely in English.
- ☐ Yes, the site is mostly in English.
- ☐ No, most or all of the site is in another language.

If you answered "No" to either of the previous two questions, please skip to the end of the form and click "Submit." Otherwise, please continue.

Which of the following aspects of politics are raised on this site?

(Please check all that apply.)

- ☐ Polls and public opinion
- ☐ Elections and political campaigns
- ☐ Legislation and law-making
- ☐ Implementation and execution of public policy
- ☐ Decisions by courts
- ☐ The actions, personality, and character of political figures
- ☐ The actions, positions, and ideologies of political parties
- ☐ Philosophical discussion about the role of government in society
- ☐ State and local politics (e.g. city government, school boards, local law enforcement)
- ☐ International relations and foreign policy
- ☐ Politics within or between countries other than the U.S.
- ☐ Other:

Overall, is politics an important topic on this site?

(Unless you checked at least one of the boxes above, this site is almost certainly not political!)

- ☐ Yes, politics is the main topic on this site.
- ☐ Yes, politics is one of several important topics on this site.
- ☐ No, politics is rarely or never mentioned on this site.

How confident are you that another person would agree with this "yes" or "no" answer?

It's very common for reasonable people to disagree on the exact definition of political content. Even if you read a site carefully, it may fall into a "gray area" where others disagree with your answer.

0 1 2 3 4 5 6 7 8 9 10

Not confident at all -- it's a
toss up whether others
would agree with my
answer.

☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐

Very confident -- just
about anyone would
agree with my answer.

About what percent of the content on this site is political?

0 1 2 3 4 5 6 7 8 9 10

0% ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ 100%

Submit

Powered by [Google Docs](#)

[Report Abuse](#) - [Terms of Service](#) - [Additional Terms](#)

Appendix C: The 200 most-linked political sites in no particular order

- [1] www.politico.com
- [2] peribadirasulullah.wordpress.com
- [3] www.npr.org
- [4] markwadsworth.blogspot.com
- [5] minx.cc
- [6] www.gaypatriot.net
- [7] abcnews.go.com
- [8] cathcon.blogspot.com
- [9] martinklasch.blogspot.com
- [10] babajidesalu.wordpress.com
- [11] likethedew.com
- [12] www.qando.net
- [13] jowlindsay.blogspot.com
- [14] upge.wn.com
- [15] blacknright.wordpress.com
- [16] www.haaretz.com
- [17] www.americanthinker.com
- [18] carloz.newsvine.com
- [19] palatialliving.blogspot.com
- [20] krishna109.newsvine.com
- [21] www.townhall.com
- [22] www.dailykos.com
- [23] www.americablog.com
- [24] www.washingtonpost.com
- [25] experimentaletc.blogspot.com
- [26] www.delawareliberal.net
- [27] artodyssey1.blogspot.com
- [28] www.bookninja.com
- [29] archive.wn.com
- [30] townhall.com
- [31] mediamatters.org
- [32] dialogic.blogspot.com
- [33] www.vivafashionblog.com
- [34] www.alanknox.net
- [35] www.telegraph.co.uk
- [36] lxa.wordpress.com
- [37] loveisntenough.com
- [38] wnbusiness.com
- [39] www.bookwormroom.com
- [40] www.slate.com
- [41] jasapaal.blogspot.com
- [42] www.wn.com
- [43] hamsters-wheel.blogspot.com
- [44] bounguine.blogspot.com
- [45] lady-light.blogspot.com
- [46] wn.com
- [47] forthardknox.com
- [48] www.blogginboutbooks.com
- [49] www.bloodhoundrealty.com
- [50] www.dumpbachmann.com
- [51] curmudgeonlyskeptical.blogspot.com
- [52] der-likedeeler.blogspot.com
- [53] auntjemimasrevenge.blogspot.com
- [54] www.nationalreview.com
- [101] bloodhoundrealty.com
- [102] takeourcountryback-snooper.blogspot.com
- [103] sarah-palin-2012.blogspot.com
- [104] smalltalkwitht.blogspot.com
- [105] bishopalan.blogspot.com
- [106] tclblogger.blogspot.com
- [107] duderay.blogspot.com
- [108] turretinfan.blogspot.com
- [109] bldgblog.blogspot.com
- [110] visnews-en.blogspot.com
- [111] agoodchoice.blogspot.com
- [112] thinkprogress.org
- [113] www.hinessight.com
- [114] uselesstriviaandmindlessrants.blogspot.com
- [115] loveandgarbage.wordpress.com
- [116] voices.washingtonpost.com
- [117] lookingforavoice.blogspot.com
- [118] dprogram.net
- [119] thegirlfromtheghetto.wordpress.com
- [120] hotair.com
- [121] grayee.blogspot.com
- [122] www2.macleans.ca
- [123] tundratabloid.blogspot.com
- [124] arbroath.blogspot.com
- [125] blog.coturnix.org
- [126] www.ft.com
- [127] www.dmzhawaii.org
- [128] edubloggerdir.blogspot.com
- [129] aw2w.blogspot.com
- [130] maggiesmetawatershed.blogspot.com
- [131] mediaconference.com
- [132] redactednews.blogspot.com
- [133] ipdragon.blogspot.com
- [134] daisysdeadair.blogspot.com
- [135] latimesblogs.latimes.com
- [136] religionnerd.com
- [137] radiograffitionline.com
- [138] blogger.xs4all.nl
- [139] www.realclearpolitics.com
- [140] lovable-liberal.blogspot.com
- [141] www.reuters.com
- [142] mauricepinayblog.wordpress.com
- [143] pajamasmedia.com
- [144] static.technorati.com
- [145] www.wordonemploymentlaw.com
- [146] catholicanalysis.blogspot.com
- [147] railwayeye.blogspot.com
- [148] jerrycaesar.blogspot.com
- [149] stupidd.blogspot.com
- [150] thecampofthesaints.wordpress.com
- [151] martynemko.blogspot.com
- [152] www.humanevents.com
- [153] www.somanyrecordssolittletime.com
- [154] news.yahoo.com

[55] myinvestingnotes.blogspot.com	[155] shigekuni.wordpress.com
[56] www.redstate.com	[156] thebluescollective.blogspot.com
[57] adelaidegreenporridgecafe.blogspot.com	[157] www.thehousenextdooronline.com
[58] shakespearessister.blogspot.com	[158] chicago.cbslocal.com
[59] www.cornwallalliance.org	[159] rosemarysutcliff.wordpress.com
[60] www.huffingtonpost.com	[160] www.salon.com
[61] cosseyedcyclops.blogspot.com	[161] pureland.blogspot.com
[62] hurricane-katrina.org	[162] acheteretentretenirsatronconneuse.com
[63] gay-persons-of-color.blogspot.com	[163] thehill.com
[64] blogfromonhigh.blogspot.com	[164] nominister.blogspot.com
[65] sotho.blogsome.com	[165] www.hqhiphop.net
[66] inhabitat.com	[166] phourdythrea.blogspot.com
[67] illuminate.newsvine.com	[167] lawhawk.blogspot.com
[68] www.washingtonblade.com	[168] www.jpost.com
[69] consul-at-arms2.blogspot.com	[169] blogs.telegraph.co.uk
[70] joninbetween.blogspot.com	[170] afrikaner-genocide-achives.blogspot.com
[71] delhistar.com	[171] www.whitehouse.gov
[72] ramblingrhodes.mu.nu	[172] www.cbsnews.com
[73] www.nytimes.com	[173] rodonline.typepad.com
[74] wefuckinglovemusic.blogspot.com	[174] www.rosecityreader.com
[75] deansoffice.blogspot.com	[175] www.thejidf.org
[76] www.freedomscost.net	[176] brutalwomen.blogspot.com
[77] blog.muzophile.com	[177] www.antifascistencyclopedia.com
[78] www.politikditto.com	[178] socraticgadfly.blogspot.com
[79] hassers.blogspot.com	[179] michellemalkin.com
[80] wnenergy.com	[180] www.bradblog.com
[81] paranormalromanticsuspensereviews.bl	[181] chrisinsouthkorea.blogspot.com
[82] zennie2005.blogspot.com	[182] www.theatlantic.com
[83] rupeenews.com	[183] www.foxnews.com
[84] paullevinson.blogspot.com	[184] www.sbnation.com
[85] www.einnews.com	[185] www.realzionistnews.com
[86] www.foreignpolicy.com	[186] mars2earth.blogspot.com
[87] cafepacific.blogspot.com	[187] www.fogcityjournal.com
[88] electronicvillage.blogspot.com	[188] boyculture.typepad.com
[89] www.perisik-rakyat.com	[189] helloskyblu.blogspot.com
[90] johnny2k-time2escape.blogspot.com	[190] mbway.blogspot.com
[91] archivelinks.wn.com	[191] capitalfax.blogspot.com
[92] schools.collegedegrees.com	[192] www.newsfortvmajors.com
[93] nethspace.blogspot.com	[193] david-mcmahon.blogspot.com
[94] news.google.com	[194] biggovernment.com
[95] sluggisha.blogspot.com	[195] news.bbc.co.uk
[96] www.davidmixner.com	[196] www.technorati.com
[97] www.lolagetslife.com	[197] technorati.com
[98] news-updatons.blogspot.com	[198] biased-bbc.blogspot.com
[99] stirredstraightup.blogspot.com	[199] climateprogress.org
[100] stevethepenguin.blogspot.com	[200] www.socialistparty.org.uk

I am looking for contact information for bloggers, so that I can conduct a research survey.

Please follow this link to answer the questions below. Answers will be screened carefully for consistency.

[http://\\${site}](http://${site})

First, we need to screen out sites that don't load, aren't in English, aren't blogs, or aren't active.

Does the site *load* properly?

- ☐ Yes, the site loads and I can read its contents.
☐ No, the site fails to load.

Is the site primarily in *English*?

- ☐ Yes, the site is entirely in English.
☐ Yes, the site is mostly in English.
☐ No, most or all of the site is in another language.

Is this site formatted as a *blog*?

(A blog is a site where the main content is a series of posts in reverse-chronological order.)

- ☐ Yes, the main content on this site is a series of posts in reverse-chronological order.
☐ No, the site follows some other format.

Has this site been updated with *new content* in the last six months?

- ☐ Yes, within the last day.
☐ Yes, within the last week.
☐ Yes, within the last month.
☐ Yes, within the last six months.
☐ No, the last update was more than six months ago.
☐ No, this site does not look like it is ever updated with new content.

If you answered "Yes" to *all* of the questions above, please continue. Otherwise, skip to the end of the form.

What is the *name* of this blog?

Is *politics* an important topic on this web site?

(By "political content" I mean content about elections, public opinion, public policy, public officials, and so on. Personal ideas, opinions be political, as long as the government is involved.)

- ☐ Yes, politics is the main topic on this site.
☐ Yes, politics is one of several important topics on this site.
☐ No, politics is rarely or never mentioned on this site.

Is politics *in the United States* an important topic on this web site?

(By politics in the U.S., I mean national, state, and/or local politics, along with U.S. foreign policy. So blog posts about a city council, diplomatic negotiations by the U.S. would all count as U.S. politics.)

- ☐ Yes, U.S. politics is the main topic on this site.
☐ Yes, U.S. politics is one of several important topics on this site.
☐ No, politics is rarely or never mentioned on this site.

How many *authors* does this blog have?

- ☐ One author
☐ Two authors
☐ Three authors
☐ Four authors
☐ Five or more authors

We need to collect *names and contact information* for as many of those bloggers as possible.

Please look carefully for contact information.

	First name	Last name	Pseudonym or made-up name	Email address
1	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>
2	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>
3	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>
4	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>
5	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>

If no email addresses are available, is there some other way of getting in touch with the author(s) of the b

(Please check all that apply.)

- ☐ Yes, there is a drop box for sending private messages to the blogger(s).
☐ Yes, the blog allows readers to post comments.
☐ Yes, the blogger lists a twitter feed.
☐ Other:

Notes/Comments:

Thank you!

I appreciate your help -- there's no way this project could happen without the help of mTurks like you.