

Measuring Media Diversity Online and Offline: Evidence from Political Websites*

Matthew Hindman[†] Kostas Tsioutsoulis[‡] Judy A. Johnson[§]

September 17, 2004

Abstract

As policymakers and scholars debate the impact of the Web on democratic discourse, one crucial question has remained unanswered: out of the millions of politically-relevant sites online, how many do citizens actually use?

To gather evidence on this question, we examine the link structure surrounding political Websites. We show that the number of hyperlinks a site receives is highly correlated with site traffic and is a key determinant of search engine ranking. In the first large-scale survey of online political content, we download and classify millions of Web pages. Links within political communities follow highly-concentrated power law distributions focused on a few hyper-successful sites.

The level of concentration we find in links and site traffic—both overall and within political communities—is comparable to or even greater than that found in traditional media. This fact challenges much of the conventional wisdom about the Internet’s influence on political life.

*We gratefully acknowledge the contributions of Lada Adamic, Larry Bartels, Adam Berinsky, Kenneth Cukier, Paul DiMaggio, Eszter Hargittai, Jennifer Hochschild, Chris Karpowitz, Gabriel Lenz, Arthur Lupia, Christopher J. Mackie, Eli Noam, Clay Shirky, and Adam Simon. Any errors are our own.

[†]Assistant Professor, Political Science Department, Arizona State University. (ASU Box 873902, Tempe, AZ 85287-3902; mhindman@princeton.edu).

[‡]Research Staff Member, NEC Laboratories America (4 Independence Way, Princeton, NJ 08540; kt@nec-labs.com)

[§]Research Staff Member, NEC Laboratories America (4 Independence Way, Princeton, NJ 08540; jaj@nec-labs.com)

1 Introduction

Social scientists and policymakers have speculated much in the past decade about the World Wide Web's impact on democratic politics—and yet they have almost never considered the link structure of the medium. This omission is a critical one. Hyperlinks are the defining feature of the Web, the strands that, collectively, weave the Web together. The interlocking patterns hyperlinks form are the reason the medium was named “the Web” in the first place.

Consequently, the link structure of the Web contains an enormous quantity of useful information. Most users see a tangible demonstration of this fact on a daily basis: PageRank, the ranking algorithm which powers the Google search engine, relies largely on the link structure of the Web to order its results. Other search engines, such as Yahoo, also pay a great deal of attention to link structure.

In this paper, we argue at length that the link structure of the Web can help us approximate the relative visibility, and the relative traffic, of political Web sites, even in the absence of cross-sectional data about the sites surfers visit. The reason for this is simple: as we show, the number of links pointing to a site is highly correlated with both its ranking in search engines, and the number of visitors that the site ultimately receives. The number of links that a site accumulates is evidence of both which sites matter for mass politics, and how much relative weight each site has earned.

The link topology of the Internet thus allows us to draw a rough map of how the attention of citizens is distributed across different sources of online information. Using cutting-edge techniques borrowed from computer science, we explore millions of Web pages, looking at topical clusters of Web sites focused on a wide variety of subjects: congress, general politics, abortion, the presidency, the death penalty, and gun control. In every case, the distribution of links within each community of sites follows a power law, where a small set of hyper-successful sites receives the vast preponderance of the links.

Using both our data on links, and link and traffic data from other sources, we then compare patterns of online concentration with those found in radio, television, and print media. We find that the winners-take-all pattern that defines online content seems to be at least as great as that found in more traditional media.

Why does this matter? Because most debates about the Web's impact are at root debates about online audience concentration. Scholars have reached very different conclusions about the Web's po-

litical influence. Some scholars have focused on the Web’s ostensible ability to alter the economy of political information—by allowing anyone to post political messages on a medium with worldwide reach, and by lowering the cost of retrieving that political information, the claim has been that the Web would alternately spur greater political engagement or balkanize public discourse. Others have argued that the digital divide, Americans’ general lack of interest in politics, and the movement of traditional media outlets and interest groups online undermine the transformative power of the new medium.

The data we present on a winners-take-all Web challenges the foundational assumptions of both of these lines of scholarship. Claims that the Web would balkanize politics, or that it would inspire greater civic engagement, have rested on the notion that the Web would steer citizens toward a more diverse set of sources for political information. Our data suggests that this assumption is precisely backwards. Both over the entire Web and within defined communities of political sites, lowering formal barriers to entry seems to have been accompanied by concentration rather than dispersion.

At the same time, data on the structure of the Web challenges those who have been skeptical that the Web will transform mass politics. The patterns we observe suggest that skepticism is justified, though on different and ultimately more fundamental grounds than previous scholars have acknowledged. The subtext of much research on politics and the Web is “if only”—if only traditional media outlets had not made the jump online, or if only all users had equal access to the Internet and greater information literacy, or if only more citizens had the motivation to participate in politics, *then* the Web would succeed in “democratizing” the flow of political information. If not for these limiting factors, citizens would be consume political information from a much broader set of sources.

Our research suggests that none of this scholarship gets at the central point. We submit that, even if one could wave a magic wand to fulfill all of the above conditions, the hierarchical structure of links on the Web would continue to sharply limit what citizens see. Some scholars have made much of the open technical standards which run the ‘Net, the “end-to-end” architecture which allows each computer online to connect to any other. But it is the link architecture of the Web which is really important—the links that provide paths to surfers, that underly search engine rankings, and that ultimately funnel citizens to a small set of winning sites within each online niche. As the Web becomes an increasingly important channel for political messages, political scientists must understand how link

structure powerfully influences how citizens interact with the Web.

1.1 Article Structure

The argument of this paper is presented in four main parts. First of all, we lay out the stakes of this research in greater detail. In a discussion intended to be suggestive rather than comprehensive, we look at scholarly debates about the Internet’s political impact. There have been longstanding arguments about whether the Internet will shift citizens’ attention away from a small set of traditional news outlets, interest groups, and opportunities for self-expression. Measuring the concentration of online audiences provides an important test of this assumption.

Second, we explain why the number of links to a site is, in the aggregate, a good proxy for the relative visibility of Web sites. There are only two ways to find new content online: users can either surf to new content by following links, or they can use a search tool like the Google search engine. This section explains that both methods direct Web traffic to sites with the greatest number of inbound hyperlinks. We present data showing that, unsurprisingly, the number of links pointing to a site proves to be highly correlated with the number of visitors the site receives.

Third, we measure the distribution of links within a diverse set of political communities. The methodology we use is innovative, involving both automated “robots” to crawl and download millions of Web pages, and highly-reliable learning algorithms to classify Web content. This analysis shows that political communities on the Web are highly concentrated: in every case a relatively small set of sites receives most of the links.

Fourth, we compare the highly-concentrated link structures we find in online political communities, along with previously-available data on Web traffic, to concentration patterns in traditional media. While we deliberately defer difficult and subjective questions about whether the diversity we find in online communities is “enough,” in relative terms the answer is clear: viewed from a national level, the concentration we find online is at least equal to that found in most traditional media.

While our discussion of the normative political consequences of the structures we find is by necessity exploratory and brief, we suggest that the results should be mixed. The good news, from the perspective of democratic theory, is that highly concentrated communities are easier for most citizens to navigate. If our central concern is to enable the least-sophisticated citizens to find relatively high-quality political

information, we could do worse than directing citizens overwhelmingly to a few mainstream sites.

The larger and unmistakable conclusion, however, is that the Web is far less open than scholars have hoped or feared. The popular wisdom that the Web is a “narrowcasting” or “pointcasting” medium may thus be misguided; our data suggests that Web audiences are just as concentrated as those for print and broadcast media. For those scholars concerned about Web-fueled balkanization, this new understanding should assuage their fears. But for those who believe that narrowcasting will transform politics, the link structure and traffic patterns on the Internet defy these expectations.

Ultimately, this research provides more questions than answers. The link structure of the Web shows us that the medium is not as we first imagined it to be. Thinking through the implications of this new data, and reconciling it with previous scholarship, will be a continuing task.

2 Scholars, Policymakers, and Online Concentration

Before we present new data on concentration within online political communities, we must explain why concentration is something scholars should be concerned with in the first place. Why should we care if the structure of the Web surrounding an online political community sends citizens to 2 sites, 20 sites, or 200 sites?

The answer is simple: most debates about the Web’s impact on politics are, at bottom, debates about audience concentration. For many political scientists—not to mention a few high-ranking public officials—the lesson of the Internet is that broadcasting’s days are numbered. The Web’s potential to decentralize where citizens get their political information is the reason they became interested in the medium in the first place.

Centralization on the Web has been a crucial point of dispute in previous scholarship, with some arguing that the Web should be markedly more decentralized than older media, and others assuming that concentration patterns should be comparable to those in the offline world. The third possibility—that that the Web might be *more* concentrated than traditional media—has rarely been considered.

2.1 Narrowcasting and Media Diversity

The importance of debates over online concentration has played out not just in the academy, but in the most visible arenas of public policy. On June 2, 2003, the Federal Communications Commission voted to

loosen restrictions on media ownership which dated back to the early 1940s. Asked about the reasoning behind this change, commission chairman Michael K. Powell explained that the Internet represented a seismic shift in the contemporary mediascape: “What’s happening now is that technology creates many different platforms and means of distributing news and content in a way that’s more dynamic and diverse, as opposed to [a time] when I say to my kids, ‘Sit down at 7 p.m., turn on Walter Cronkite, we’ll get our news and go to bed’ ” (Manjoo 2003).¹

The controversy surrounding the FCC decision caused some to question the commission’s real motives. But for those inclined to take a cynical view of the proceedings, one should remember that mainstream political scientists have long made far more extreme statements about how the Internet changes the media environment for citizens. Actual data on the structure of the Web can provide a test of those assumptions, and help revise our understanding of the Web’s most basic characteristics.

For our purposes, it is useful to divide scholarship on the Web into three camps: cyberoptimists, who think that the Web will change politics for the better; cyberpessimists, who think that the Web will change politics for the worse; and cyberskeptics, who argue that for politics the Web will not change much of anything.²

The optimists and pessimists disagree about many things, but both schools have argued repeatedly that the Internet is not a mass medium. In their view, as in Chairman Powell’s, the Web represents a fundamental shift from broadcasting to “narrowcasting,” where content is not produced for the general public but tailored for a much smaller audience. Cass Sunstein believes that we are fast approaching a future where “Technology has greatly increased people’s ability to ‘filter’ what they read, see and hear. General interest newspapers and magazines are largely a thing of the past. The same is true of broadcasters” (Sunstein 2001, p. 3). With citizens getting their political information from countless unreliable and polarizing sources, Sunstein argues, the Web means that democratic discourse will become a balkanized mess. In the same vein, Andrew Shapiro worries that online echo chambers of like-minded citizens will undermine public discourse and make collective decisionmaking impossible (Shapiro 1999). Joseph Nye suggests that “the demise of broadcasting and the rise of narrowcasting may fragment the sense of community and legitimacy that underpins central governments” (Karmark

¹Public outcry caused the most important parts of the FCC decision to be rescinded by the U.S. Congress. Yet notions about the Internet’s openness continue to be a key justification of the FCC’s policies. In the wake of key court decisions and the newly-announced policy, the FCC created a formula to measure and formalize the level of diversity within a media market. The FCC’s formula places a large weight on the Internet as a source of diversity.

²This taxonomy follows to Norris 2001.

and Nye 2002, p. 10).

Though other scholars are more hopeful about the future of democratic practice, they often agree that the Web will lead citizens to a greater diversity of political information. As Lupia and Sin suggest,

The World Wide Web [...] allows individuals—even children—to post, at minimal cost, messages and images that can be viewed instantly by global audiences. It is worth remembering that as recently as the early 1990’s, such actions were impossible for all but a few world leaders, public figures, and entertainment companies—and even for them only at select moments. Now many people take such abilities for granted (Lupia and Sin 2003).

Lupia and Sin’s arguments about collective action rely on the assumption that, thanks to the Internet, the attention of the public will be more dispersed. After all, the ability of citizens to post content online means little if that content is not seen by others.

In the political science literature, concerns about balkanization are closely connected to an older set of worries about the personalization of information preferences.³ Ithiel de Sola Pool argued in the early 1990’s that excessive personalization of media content was a pressing problem for democratic theory, and one that required creative policy solutions (de Sola Pool 1993). Lance Bennett concurred, arguing that the subsequent emergence of the Internet as a mass medium fulfilled much of what de Sola Pool had prophesied. Declared Bennett, “The capacity to submit and refine personal information preferences already exists in the search engines and information sites of the Internet. Future research and policy analysis must address how ever more personalized preferences can be accommodated in collective decision making” (Bennett 1998, p. 757).

In other ways, too, audience concentration directly impacts theories of political participation in the digital age. One longstanding hope has been that information technology would help overcome geographic barriers and solve the Madisonian dilemma of deliberation and participation in an extended republic (Dahl 1989, Barber 1984). However, a highly concentrated Web would likely make it hard for all but a few “ordinary citizens” to post their views prominently—and, conversely, to read the views of other ordinary citizens—except on a small number of prominent sites. Most political speech posted online would be hidden by countless other Web pages.

³Personalization of content includes not just diversity in the sources citizens turn to for political information, but increasing selectivity within those sources—for example, filtering the news received from a broad Web portal to focus only on a few topics particular interest. In the context of the scholarship described here, however, source diversity is the most important concern.

2.2 Cyber-Skeptics

Notions that the Web would alter the political information consumption of citizens have never been universally accepted. Partly, skepticism has come out of literature on the so-called “digital divide.” There has been persistent evidence over the past decade that access to the Web follows existing social cleavages. Even as the pool of users has expanded dramatically, disadvantaged groups—blacks, Hispanics, the poor, the elderly, the undereducated, those who live in rural areas—have lagged behind in their access to and use of the Net (NTIA 2000, 2002, Bimber 2000, Wilhelm 2000). Other scholars have shown that the skills that users need to use the Web effectively are similarly stratified (Hargittai 2003, DiMaggio and Hargittai 2001, Norris 2001).

Demographics and disparities in user skill are not the only impediments that scholars have pointed to. Doris Graber argued early on that most citizens would continue to focus on a small set of outlets, because “little [political information] has been added that is genuinely new or that enriches the information supply beyond the offerings of the far smaller circle of ‘old media’ ” (Graber 1996, p. 33). Other scholars argued that the quick migration of traditional outlets onto the Web helped short-circuit the shift to a broader, more diverse set of political sources (Davis 1998, Bimber 2003, Davis and Owen 1998). The default assumption in this school of scholarship has been that the online media environment should closely resemble the offline one, and that concentration on the Web should follow the existing patterns seen in print and broadcasting.

These same scholars, noting the movement of traditional media online, have also unanimously raised questions about the political motivations of users. Survey data consistently shows that most Americans pay little attention to politics. Norris labels this divide between the engaged and the disaffected the “democracy gap,” and suggests that it works to constrain the number and diversity of political sites that citizens use (Norris 2001).

For scholars who think that the Web will dramatically alter politics, and for those who do not, audience concentration online has thus been a central issue. Everyone agrees that the Internet provides citizens with an unprecedented array of politically relevant information. Yet the question remains: with millions of new information sources at their disposal, how many will citizens actually use? Is the information the public consumes online as diverse in practice as it is supposed to be in theory? How far have we really come from the Cronkite era? The link structure of the Web can help us evaluate

this questions, even in communities which are too small or too socially unpopular to study with cross-sectional data.

3 What Link Structure Can Tell Political Scientists

The previous section discusses the ways in which scholarly debates about the Web's impact on politics depend critically on assumptions about the concentration of online audiences. We now explain why the number of links pointing to a site provides a good, if rough, measure of the relative popularity of political Websites.

We begin, first, by reviewing previous scholarship on link structure and traffic patterns for the Web as a whole. Second, we discuss search engine algorithms and patterns of Web surfing. Using a search engine, or surfing away from known sites, are in practice the only ways that users can find new content. We show that both methods funnel users to the sites that have the greatest number of inbound hyperlinks. Finally, to confirm these theoretical conclusions, we present data on the relationship between the number of links pointing to a site and the number of visitors that site receives. As we expect, the correlation is high.

3.1 The Link Structure of the Web

The structure of the Web has been a remarkably fertile area of scholarship in recent years. Though most of this work has been done by computer scientists and applied physicists, the striking patterns they have found in the apparent chaos of the Web should give political scientists cause to rethink the Web's political implications.

In looking at the structure of the Web, the central finding is that links between sites obey very strong statistical regularities. Over the entire Web, the distribution of both inbound and outbound hyperlinks follows a power law or scale-free distribution (Barabasi and Albert 1999, Kumar et al. 1999). More precisely, the probability that a randomly selected Web page has K links is proportional to $K^{-\alpha}$ for large K .

Data follow a power law distribution when the size of an observation is inversely and exponentially proportional to its frequency. For example, the distribution of wealth, as Pareto famously explained, is a power law distribution, where 20% of the population controls 80% of the wealth (Pareto 1897).

Numerous other social and natural phenomena follow this pattern as well, from earthquakes to intracellular protein networks, from the size of firms to the size of cities, from the severity of wars to the number of sexual contacts (Huberman 2001, Krugman 1994, Cederman 2003, Liljeros et al. 2001).

As the diverse scholarship related to power laws demonstrates, power law structures can be generated by very different underlying processes. But in every case, a power law distribution leads to remarkably inegalitarian outcomes. To get a sense of just how extreme these results can be in practice, imagine a hypothetical community where wealth is power-law distributed. At one end of the spectrum, there is one millionaire, ten individuals worth at least 100 thousand dollars, a hundred people worth 10 thousand dollars, and a thousand people worth at least a thousand dollars. At the opposite end of the spectrum, 1,000,000 people have a net worth of \$1. In this hypothetical community, wealth is distributed in proportion to the function $K^{-\alpha}$, where $\alpha = 1$.

In the context of the Web, studies have found the online environment to be far more concentrated even than the hypothetical example above, generating values of $\alpha \approx 2.1$ for inbound hyperlinks, and $\alpha \approx 2.72$ for outbound hyperlinks (Kumar et al. 1999, Barabasi et al. 2000, Lawrence and Giles 1998, Faloutsos, Faloutsos and Faloutsos 1999).⁴ A few popular sites (such as Yahoo or AOL or Google) receive a large portion of the total links; less successful sites (such as most personal Web pages) receive hardly any links at all.

That's not all. Traffic, like link structure, follows a power-law distribution with roughly the same parameters (Huberman et al. 1998, Adamic and Huberman 2000). There is thus a small set of sites that receive most of the links, and a small set of sites that receive most online visitors. For the purposes of this paper, it is important to demonstrate that these two groups are in fact one and the same.

We do this in two ways. In the next sections, we explain *why* the number of links pointing to a site is such a powerful predictor of traffic: both surfing patterns and search engines send users overwhelming to the sites that have accumulated the most links. Then, we test this theoretical expectation in practice, examining data on the correlation between links and site traffic.

⁴Barabasi et al. and Kumar et al. seem to disagree on the value of α for outgoing hyperlinks; Barabasi et al. propose a value of $\alpha = 2.4$. This scholarship also shows that these parameters have been highly stable over time, even as the Web has undergone explosive growth.

3.2 Finding Online Information

Before one can visit a Web site, one must be able to find that site in the first place. Sites that are already known can be visited just by typing in the URL or by using a bookmark within a Web browser. Content the user has *not* previously seen, however, can be found in only two ways. First, it can be discovered by surfing away from known sites; or second, it can be found with the help of online search tools such as Google or the Yahoo directory service. In both cases, the number of inbound hyperlinks turns out to be a crucial determinant of a Web page's visibility.

Much of the association between inbound links and traffic is simple: hyperlinks exist to be followed. The more hyperlinks there are to a given site, the more chances users on connecting sites have to follow them. In the aggregate, more paths to a site means more traffic.

What is true for the behavior of individual surfers is doubly so for search engines. Even when the underlying ranking algorithm is far more complex, search engines in practice rank sites by the number of inlinks they receive. With search engines, too, users end up at sites with lots of links.

To understand why this is the case, consider the recent history of the Google search engine. The first generation of search engines, such as Alta Vista, focused on keyword density and other characteristics found within individual Web pages. Google's contribution was to take a broader view, and use the connections *between* different Web sites to find the best content. Brin and Page developed PageRank, a recursive algorithm in which sites which receive lots of links, from *other* sites that receive lots of links, are the ones ranked most highly (Brin and Page 1998, Pandurangan, Raghavan and Upfal 2002). In essence, sites are ranked in a popularity contest, in which each link is a vote, and the votes of popular sites carry more weight.⁵

Both surfing away from known sites and the use of search tools thus privilege the same set of Web pages. Sites which are heavily linked to by other sites become prominent; most other sites are likely to be ignored.

According to Nielsen/Netratings, the Google and Yahoo search engines together control more than 95 percent of the search engine market.⁶ One might think that a less concentrated search engine market

⁵As time has passed, Google has increasingly incorporated other factors into its rating algorithm. Though these refinements make it harder to manipulate search engine results, they make only modest changes in the overall rankings—particularly in the first few pages of search results. Even today, PageRank and other measures of link structure continue to be the backbone of Google's ranking system.

⁶February 2004 data show that the Google engine powered 49.7 percent of U.S. searches, including AOL searches performed using Google's licensed technology. The Yahoo engine provided 45.4 percent of the search results for the U.S.

would help ensure diversity in the content seen. But in truth, the popularity contest dynamics we see with PageRank are difficult to avoid. The HITS algorithm is one widely-known alternative to PageRank, and uses the mutually reinforcing structure of “hubs” and “authorities” to rank results (Kleinberg 1999, Marendy 2001). But Ding et al. show that despite the fact that the HITS approach is “at the other end of the search engine spectrum” from PageRank, it ranks the same set of sites first. Indeed, both engines—and any likely competitors—produce results that are hardly different than just ordering sites by the number of inlinks they receive (Ding et al. 2002).

This research thus explains why, though Yahoo’s algorithm is different than Google’s, in practice it produces nearly identical results. No matter which search engine is used, the small number of sites with large numbers of inbound hyperlinks are returned first.

3.3 The Relation Between Inbound Links and Web Traffic

To recap: we know that over the entire Web both traffic and links are power-law distributed. We also have strong theoretical reasons to believe that traffic will be driven to heavily-linked sites. But how close is the relationship between link structure and site visits in practice?

Both our own analysis and that of other researchers suggests that, in the aggregate, the link is reasonably strong. First, Dr. Lada Adamic of Hewlett Packard Laboratories provided us with on data on links to Web sites along with the number of visitors these sites receive. The site visit data are from a randomly-selected, anonymized set of users from a large Internet service provider. They include visits by 60,000 users to 120,000 sites; the link data for visited sites is provided by Alexa.

In this data, the number of hyperlinks pointing to a site and the number of visits it receives are highly correlated, generating a correlation coefficient of .704. The raw number of hyperlinks pointing to a site does seem to a good predictor of its traffic.

Recent data also shows that the connection between links and traffic is equally strong within defined subcommunities of the Web. In the past few years the Weblog community has become an important part of online political discourse. Several sites track the number of links that each of these online journals receive from other Weblogs. Importantly, a large number of Weblogs also use `sitemeter.com` to track visitors, providing a consistent comparative measure of site traffic.

market, including results for Microsoft’s MSN.com portal site (Nielsen-Netratings 2004).

Using this data, researchers have shown that links and traffic have roughly the same level of correlation within Weblogs as in the earlier data on the Web as a whole (Shirky 2004). As we might expect with this power law data, links are best at predicting the traffic of popular sites.

The number of inbound hyperlinks to a Web site is strongly associated with its overall traffic. But do political sites on the Web follow a power-law distribution? While the global properties of the Internet are quite clear, subgroups of sites seem to diverge quite significantly from the overall pattern. Within specific categories of sites—for example, sites for publicly listed companies, university homepages, and newspaper homepages—researchers have found that the distribution of hyperlinks obeys a less extreme, roughly log-normal distribution (Pennock et al. 2002). However, these communities that have been found to deviate from the expected distribution have done so to widely differing degrees. It is unclear whether we should expect subcategories of political sites to be among them.

The ultimate conclusion is that the Web’s ability to present a broad range of sources for political information depends on the link structure found among subgroups of political Web sites. Still, the only way to understand the extent and structure of political information online is to measure it directly. The next section proposes methodology to do exactly that.

4 The Link Structure of Online Political Communities

We have explained both why online audience concentration is a central concern for political scientists, and why the number of links pointing to a site can serve as a proxy for the number of visitors that site receives. The next step is to put this understanding to use, by measuring the link structure surrounding political Web sites. The central goal is to draw a rough map of how concentrated these categories of political content look from the perspective of the average user.

4.1 Methodology: What Does the Average User See?

The methodology we use in this paper surveys the portions of the Internet that an average user is likely to encounter while looking for common types of political information. It is explicitly not an attempt to map every political site online, or even every political site in a given category. The purpose is not to overcome the limitations imposed by the scale of the Web; rather, it is to demonstrate the biases those limitations introduce in the number and types of sites encountered by typical users.

The research design we have chosen comes out of a large body of established computer science research. (Part of that research is summarized in the “Appendix on Methodology” at the end of the article.) The methodology we implement has four main parts:

1. Create 12 lists of 200 highly-ranked “seed sites” in a variety of political categories. Six categories are chosen; in each category, one list is taken from Google search engine results, and one is taken from the Yahoo directory service.
2. Build Web robots to crawl outward from these 200 sites, following every link in turn, 3 links deep. For each crawl, this requires downloading approximately 250,000 HTML pages, or approximately 3,000,000 pages total.
3. Classify these downloaded pages using Support Vector Machine (SVM) algorithms, to see whether newly encountered pages are relevant to the given category—if, for example, a page discovered by crawling away from gun control sites also focuses on gun control. Those pages that do belong in a particular category are classified as “positive.”
4. For each of the 12 crawls, analyze the distribution of inlinks within the set of “positive” sites.

In this research we compare the relative level of concentration across a diverse set of communities focusing on politics. Ultimately, six categories of Web sites were chosen: abortion, gun control, the death penalty, the U.S. congress, the presidency, and the catchall category of “general politics.”

It is clearly infeasible to classify the downloaded Web pages with human coders. Even if one could classify 120 Web sites an hour, it would take an individual working 8 hours a day 10 years to classify 3,000,000 pages. Human categorization also raises questions of bias and subjectivity.

To solve this problem, we classify these Web sites automatically using Support Vector Machines, or SVMs. (The technical operation of SVMs are described in the Appendix.) The SVM classifier produces highly reliable categorization of relevant Web pages. Most importantly, it produces very few false positives. Randomized human coding of SVM-classified sites shows that approximately 98 percent of sites in the positive set are correctly classified.

The choice of seed sites is obviously an important one. Not only does this set of sites determine the starting point for the Web crawlers, and thus the area of the Web which is ultimately crawled, these sites are also used to teach the Support Vector Machines to recognize relevant content. We

	Downloaded	Topical (SVM)	SVM unsure
Abortion (Yahoo)	222,987	10,219	717
Abortion (Google)	249,987	11,733	1,509
Death Penalty (Yahoo)	212,365	10,236	1,572
Death Penalty (Google)	236,401	10,890	938
Gun Control (Yahoo)	224,139	12,719	1,798
Gun Control (Google)	236,921	13,996	1,457
President (Yahoo)	234,339	21,936	2,714
President (Google)	272,447	16,626	3,470
U.S. Congress (Yahoo)	215,159	17,281	2,426
U.S. Congress (Google)	271,014	21,984	4,083
General Politics (Yahoo)	239,963	5,531	1,481
General Politics (Google)	341,006	39,971	10,693

Table 1: This table illustrates the size of the Web graph crawled in the course of our analysis, as well as the number of sites that the SVM classifiers categorized as positive. The first column gives the number of Web pages downloaded. Columns two and three give the number of pages which are classified by the SVM as having content closely related to the seed pages, as well as the pages about which the SVM was hesitant.

were initially concerned that there may be systematic biases between human-categorized content and the sorts of machine-categorized content returned by search engines. Therefore, in each category, we analyze both seed sets generated by Google, and seed sets taken from the human-categorized Yahoo directory. Ultimately, both the Google and Yahoo seed sets lead to the same conclusions.

For the Web as a whole, we know that approximately 75 percent of user search behavior terminates without going deeper than three links (Huberman et al. 1998; discussion in Appendix). Therefore, we are confident that these techniques capture most of the material accessible from Google or Yahoo in a given issue area.

4.2 Results

The six political topics examined are quite different from one another, and our research design introduces numerous sources of potential heterogeneity. The level of consistency in our results, therefore, is all the more striking. All twelve of the crawls reveal communities of Web sites with similar organizing principles and similar distributions of inbound hyperlinks.

First, let us examine the scope of the project. Table 1 lists the number of pages downloaded, as well as the results of the SVM classification. The size of the crawls is quite large—most weighed in at slightly less than a quarter of a million pages. We downloaded and classified nearly 3 million pages

	Yahoo	Google	Overlap
Abortion	10,219	11,733	2,784
Death Penalty	10,236	10,890	3,151
Gun Control	12,719	13,996	2,344
President	21,936	16,626	3,332
U.S. Congress	17,281	21,984	3,852
General Politics	5,531	39,971	1,816

Table 2: This table gives the overlap, on a given political topic, between the crawls generated by the Yahoo seed set and that generated with the first 200 Google results. The global overlap is significant, and closer examination of the data suggests that overlap is nearly complete for the most heavily linked pages in each category.

(not accounting for sites that might have been included in more than one of the dozen crawls). The size of the SVM positive sets seem to vary by the type of subject they examine. Seed sets focused on a particular political issue were smaller than those which focused on the presidency or the U.S. congress. Out of the large number of pages crawled, only a small fraction were relevant to the given category.

Table 1 suggests that the SVM classifier is good but not perfect. Very few sites in the negative set are misclassified, and the positive set is almost completely free of false positives. There are a significant number of sites, however, which are quite near the decision boundary drawn by the SVM, and which are thus classified as “unsure.” Sites about which the SVM was hesitant range from roughly 7 to 25% of the size of the positive set. Human coding of these sites suggests that most should be included in the positive set. Secondary analyses conducted with “unsure” sites included in the positive set found no substantive differences in the results detailed below—if anything, the results would be even stronger with their inclusion.

In several cases, the the Google and Yahoo seed sets were quite different. This was initially a source of some concern, even within our research group, that the communities crawled and the sites identified might not be directly comparable. Table 2, which shows substantial overlap between the positive sets from the different Yahoo and Google crawls, does much to alleviate those fears. It reinforces our conviction that the Yahoo and Google crawls are both exploring the same communities, and provides a clear demonstration of the small diameter of the Web. Most of the pages in the positive set are relatively obscure, and receive only a few inlinks. The least overlap occurs with pages with one hyperlink path to them. Among the most heavily linked pages, the overlap between the Yahoo and Google results is almost complete.

	SVM positive set	Links to SVM set	Within-set links
Abortion (Yahoo)	10,219	153,375	121,232
Abortion (Google)	11,733	391,894	272,403
Death Penalty (Yahoo)	10,236	431,244	199,507
Death Penalty (Google)	10,890	291,409	149,045
Gun Control (Yahoo)	12,719	274,715	178,310
Gun Control (Google)	13,996	599,960	356,740
President (Yahoo)	21,936	1,152,083	877,956
President (Google)	16,626	816,858	409,930
U.S. Congress (Yahoo)	17,281	365,578	310,485
U.S. Congress (Google)	21,984	751,306	380,907
General Politics (Yahoo)	5,531	320,526	88,006
General Politics (Google)	39,971	1,646,296	848,636

Table 3: This table gives the number of links to sites in the SVM positive set, from both outside the set and from one positive page to another. Note that, in most cases, links from other positive pages provide the majority of the links.

We have therefore seen that the collection of Web pages available to the majority of users of the most popular search tools is between 10,000 and 22,000 for all but one of the areas studied. Given the vastness of the medium, these accessible pages are likely only a fraction of all pages on the topic. Of even greater interest than the size of these topical communities, however, is the way in which they are organized. Table 3 gives an overview of the link structure leading to these relevant pages.

Globally, the Web graph is quite sparse; a randomly selected series of pages will have few links in common. In contrast, here we find that the number of links between these positive pages is uniformly large. Even more telling, for 10 of the 12 crawls, links from one positive page to another account for more than half the total. This fact increases our confidence that we have identified coherent communities of pages.⁷

Ultimately, however, what we want to know is the distribution of these inbound links. The first column of Table 4 contains the number of *sites* in each category which contain at least one positive *page*. For example, `abortionfacts.org` is a prominent anti-abortion Web site. `Abortionfacts.org` contains within it many Web pages that are relevant to the abortion debate. If what we are interested in, however, is the number of sources of political information, it makes greater sense to count all of the

⁷It is worth noting that the results shown are based on raw data, and may thus inflate somewhat the connectedness of the graph. To take one example: `moratoriumcampaign.org`, a popular site opposed to the death penalty, contains a number of heavily cross-linked relevant pages—and relevant page *A* may even contain more than one link to relevant page *B*. Eliminating cross-links between pages hosted on the same site eliminates a large portion of the links. The distribution of inlinks, however, remains stubbornly power-law distributed. Because we believe that the total number of inlinks is the best predictor of a site’s visibility and traffic, our analysis focuses on the raw numbers.

	Sites	Links to top site (%)	Top 10 (%)	Top 50 (%)
Abortion (Yahoo)	706	15.4	43.2	79.5
Abortion (Google)	1,015	31.1	70.6	88.8
Death Penalty (Yahoo)	725	13.9	63.5	94.1
Death Penalty (Google)	781	15.9	53.5	88.5
Gun Control (Yahoo)	1,059	28.7	66.7	88.1
Gun Control (Google)	630	39.2	76.8	95.9
President (Yahoo)	1,163	53.0	83.2	94.9
President (Google)	1,070	21.9	65.3	90.9
U.S. Congress (Yahoo)	528	25.9	74.3	94.8
U.S. Congress (Google)	1,350	22.0	51.4	82.3
General Politics (Yahoo)	1,027	6.5	36.4	70.3
General Politics (Google)	3,243	13.0	44.0	74.0

Table 4: This table demonstrates the remarkable concentration of links that the most popular sites enjoy in each of the communities explored. The first column lists the number of sites that contain at least one positive page; note that many sites contain numerous relevant pages. Columns 2, 3, and 4 show the percentage of inlinks attached to the top site, the top 10 sites, and the top 50 sites in a given category.

pages at `abortionfacts.org` as a single unit. The number of sites offering political information must, by definition, be smaller than the total number of pages.

The most important results, however, are captured in the other three columns of Table 4. Here we find the percentage of inlinks attached to the top site, the top 10 sites, and the top 50 sites in each crawl. The overall picture shows a startling concentration of attention on a handful of hyper-successful sites. Excluding one low-end outlier, the most successful sites in these crawls receive between 14% and 54% of the total links—*all to a single source of information*.

Perhaps even more telling is the third column, which shows the percentage of inlinks attached to the top ten sites for each crawl. In 9 of the 12 cases, the top ten sites account for more than half of the total links. Across these dozen examples, the top 50 sites account for 3–10% of the total sites in their category. But in every case, the top 50 sites account for the vast majority of inbound links.

There is thus good reason to believe that communities of political sites on the Web function as winners-take-all networks. But is the inlink distribution among these sites governed by a power law? The answer seems to be yes.

To see exactly what the inter-community link structure looks like in practice, consider the figures below. Figure 1 looks at sites which contain information on the presidency; Figure 2 looks at sites devoted to the death penalty. One is generated from a Yahoo seed set; the other is from Google.

President--Yahoo

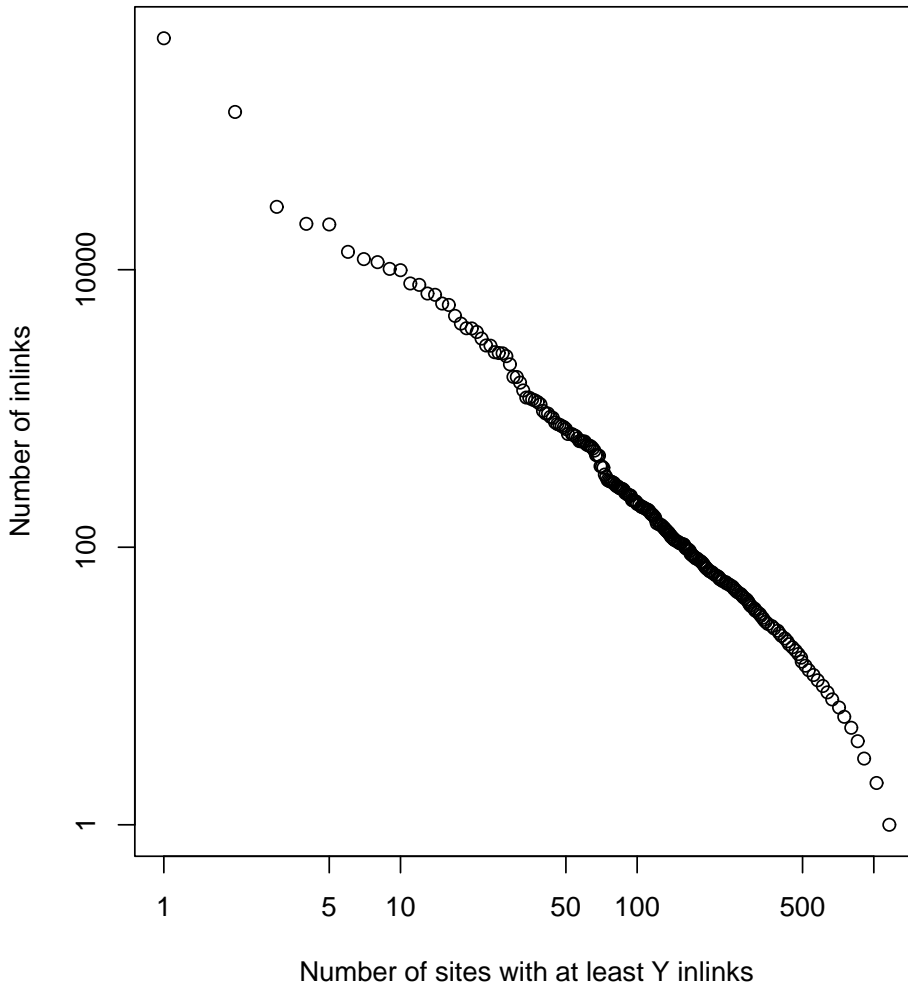


Figure 1: This chart shows the distribution of inbound hyperlinks for sites which focus on Pres. George W. Bush. Both axes are on a log scale. Note that the data form an almost perfectly straight line—unmistakable evidence of a power-law distribution.

The unmistakable signature of a power law distribution is that, on a chart where both of the axes are on a logarithmic scale, the data should form a straight line. This is precisely what Figure 1 shows—a textbook power law distribution. A similar but less exact pattern is evident in Figure 2, which is more typical of the communities crawled. Here the line formed by the data on the log-log scale bulges outward slightly; the slope of the line gets steeper as the number of sites increases. The death penalty community deviates from a power law at the tails—particularly among the set of most popular sites, where a pure power law would produce astronomical numbers of links.⁸

Overall, power laws do an excellent job of characterizing link distribution within these communities. Table 5 shows the results of fitting a power law to the data gathered by each of the 12 crawls. In this case, the model chosen is a simple ordinary least squares regression. The dependent variable is the log

⁸The slightly curvilinear shape—which forms a soft, downward-facing parabola in the log-log scale—may suggest an admixture between a power law and some other distribution with an extreme skew (such as a log-normal distribution with a mean of 0).

Death Penalty--Google

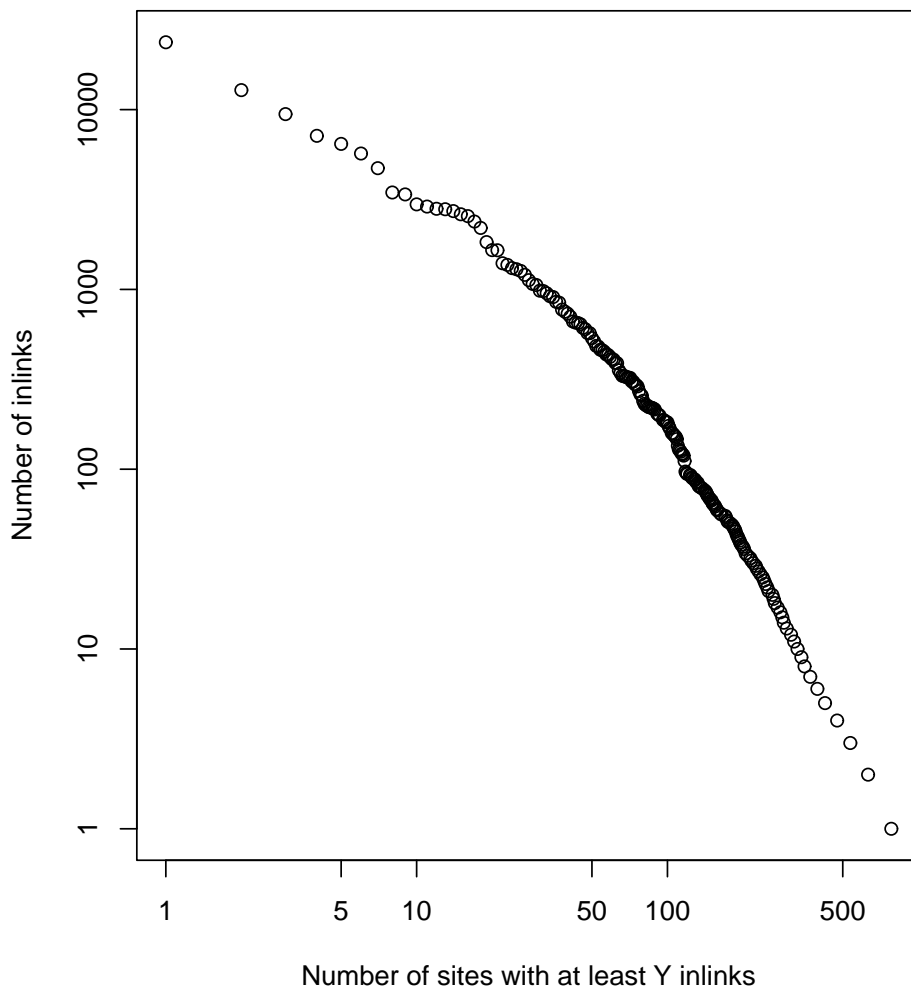


Figure 2: This figure illustrates the distribution of inlinks for sites focusing on the death penalty. Here again we see strong evidence of a power-law distribution, although there is a slight upward bulge to the plotted data. Fitting a power-law to this data produces an R^2 of .9516—the second-lowest among the communities explored.

of the number of links pointing to a given Web site. For example, if site Q has 1500 inlinks, its value on the dependent variable is equal to $\ln(1500)$, or 7.31. The explanatory variable is the log of the number of sites which have at least as many inlinks as site Q . Since a power law relationship between the two variables should produce a straight line on a log-log scale, a linear regression on the log-transformed data is a straightforward way of testing how well such a distribution fits the data. In this context, the constant is the log of the number of inlinks which the model predicts for the community's most popular Web site.

The results of this analysis show that, with a few caveats, a power law fits the distribution of inlinks within political communities quite well. The Yahoo abortion community is a markedly poorer fit than the other 11 communities explored, though the power law model still produces an R^2 of .9016. The power law model consistently predicts greater numbers of inlinks for the four or five most successful

	Coefficient ($-\alpha$)	Constant	R^2
Abortion (Yahoo)	-1.544	11.834	.902
Abortion (Google)	-1.488	11.819	.972
Death Penalty (Yahoo)	-1.684	12.007	.977
Death Penalty (Google)	-1.958	13.960	.952
Gun Control (Yahoo)	-1.458	11.650	.961
Gun Control (Google)	-1.806	13.113	.968
President (Yahoo)	-1.659	13.014	.992
President (Google)	-1.705	13.285	.975
U.S. Congress (Yahoo)	-1.909	13.239	.971
U.S. Congress (Google)	-1.530	12.952	.953
General Politics (Yahoo)	-1.252	10.583	.956
General Politics (Google)	-1.454	13.536	.977

Table 5: This table shows the results of fitting a power law to the 12 communities explored, by means of an OLS regression on the logged data. The dependent variable is the log of the number of inlinks that a given site (e.g. site Q) has received; the explanatory variable is the log of the number of sites in the sample that have at least as many inlinks as site Q . If a power law follows the form $K^{-\alpha}$, the coefficient above is equal to $-\alpha$, the slope of the power law line on a log-log scale. The constant represents the log of the number of links that the most popular site is predicted to receive.

sites than we see in the data; to a lesser degree it underpredicts the number of sites that have only a handful of links. These deviations, particularly in the upper part of the curve, are substantively significant, as they dilute the concentration of attention on the small number of successful sites.

Still, even with outliers at both tails of the distribution, the power law model produces an R^2 greater than .95 in each of the remaining communities. The body of the data, in *every* community, adheres stubbornly to a power law, and omitting the 5 highest and lowest link values usually produces a near-perfect fit. Inlink distribution within political communities is bound by powerful statistical regularities.

4.3 Site Visibility

Whether online communities are better characterized by power laws or by some other variety of extremely skewed distribution is, of course, not the central point. For political scientists concerned about the level of concentration within communities dedicated to political expression, two lessons are clear. First, the number of highly visible sites is small by any measure. It seems a general property of political communities online that a handful of sites at the top of the distribution receive more links than the rest of relevant sites put together. Second, comparative visibility drops off in a rapid and highly regular

fashion once one moves outside the core group of successful sites. Falloff in site visibility is not linear; rather, it follows an exponential function over many orders of magnitude. Given the diversity both in seed sets and in the types of communities explored, these results are surprisingly strong and consistent.

There is an often-repeated belief that the Internet is a hotbed of grass-roots political activism. In the communities that we examine, however, this belief seems to be unfounded. Almost all prominent sites are run by long-established interest groups, by government entities, by corporations, or by traditional media outlets.

There is one more point that deserves emphasis: the power law structure persists even if these sites are broken down into sub-communities. In our two crawls of the abortion community, for example, pro-choice sites outnumber pro-life sites by a margin of roughly three to one. However, both pro-life and pro-choice sites are governed by a power law. Although the slope is different across the two groups, the overall structure continues to focus attention on a few top sites. The same pattern is evident in the gun control and death penalty communities, which both contain clearly opposing subgroups. The structure of political groups on the Web thus may loosely be termed fractal in nature—portions of the community mirror the structure of the whole. This Russian nesting doll structure, dominated by power laws at every level, has important implications for politics.

5 Comparative Concentration

We began by explaining why online concentration is an issue of concern to political scientists and policy makers. We detailed why link structure is a good proxy for measuring audience concentration. And we have now looked closely at the link structure surrounding a diverse set of political Web sites. While these communities of sites do vary somewhat in their level of centralization, in almost every case they seem to follow a winners-take-all power law distribution.

Still, the Internet's political consequences emerge most clearly from a comparative perspective. How do the patterns we find in the online world—both in the aggregate and within focused political communities—compare to those that we are accustomed to in traditional media?

Our analysis places online concentration alongside television programming, newspaper and magazine circulation, and radio station listenership. None of these media are exactly comparable with the Web, or with each other—a Web site is not interchangeable with a radio station or a newspaper. Yet if

we should be cautious in interpreting the results, the larger patterns are difficult to ignore. According to two very different metrics, the concentration of audience share online seems at least as great as that found in the offline world.⁹

5.1 Data

For this analysis, we gathered a series of diverse data sets that shed light on how patterns of offline concentration compare to online media. First of all, we examine data on television ratings. We look at 171 different prime-time television programs, using public data from September 2003 available from A.C. Nielsen.

Second, we examine data on radio station listenership. We compile a national-level data set using Arbitron corporation rating data on all commercial radio stations, both AM and FM, within the U.S.'s top 50 radio markets. These 50 markets serve 128 million people age 12 or older, roughly half of the nation's total 12-and-older population. Our data set covers 1289 commercial stations.

Third, we look at data on print media, both newspapers and magazines. In both cases, nationwide circulation data comes from the Audit Bureau of Circulation. The ABC tracks the circulation of 653 magazines and 1058 newspapers nationwide.

We weigh this data on television, radio, and print against a number of different data sets on online concentration. First of all, of course, we look at a representative subsection of the 12 political communities described in the previous section. At the same time, we look at broader data sets about the Web. We look at concentration patterns of links across the entire Web, with data provided by HP Labs. We look at the universe of news Websites, with traffic data drawn from Nielsen/Netratings. And we look at the concentration of unique visitors across all Weblogs, the easiest online community to gather visitor data on, using information drawn from N.Z. Bear's Weblog Ecosystem project found at truthlaidbear.com.

All of this data is national, not regional, in scope. Radio stations in Cleveland and Baltimore cannot compete with each other for listeners, but every Web site in a given niche competes directly against all the rest. One aim of this analysis is to compare locally fragmented media against online content

⁹Yim 2003 finds that, in traditional media, concentration increases with the number of outlets available. Comparing circulation figures of the top 100 newspapers with the number of links their Web sites receive, Hamilton suggests that the economics of producing online news may result in concentration rather than dispersion (Hamilton 2004, Ch. 7).

which does not face the same geographic restrictions.

5.2 Metrics for Concentration

With these data sets in hand, we must consider what sort of measures can fruitfully be used to illustrate disparities in concentration. We use two metrics commonly found in other areas of the social sciences to provide benchmarks of concentration across mediums.¹⁰

The first of these is the Gini coefficient. Originally developed in the early 20th century to measure income inequality, the Gini coefficient can be used, as Corrado Gini himself declared, to calculate relative inequality for almost any resource (Gini 1921). The Gini coefficient is the mean difference across all observations between the Lorenz curve and the line of perfect equality.¹¹ Stated formally, if y is a vector of incomes, with extreme values of y_{min} and y_{max} , a mean of μ , and a cumulative distribution of $F(y)$, the Gini coefficient can be calculated as follows:

$$G = \frac{\int_{y_{min}}^{y_{max}} F(y)[1 - F(y)]}{\mu}$$

The Gini coefficient produces possible values between 0 and 1.

The second measure of inequality that we adapt for these purposes is the Herfindahl–Hirschman Index, or HHI. Originally developed to measure firm power within industries, HHI is calculated by taking an observation’s total resource share expressed as a percentage, squaring it, and taking the sum across all observations. Expressed more formally, the Herfindahl–Hirschman Index can be calculated as:

$$HHI = \sum_1^N P_i^2$$

where P_i is the percentage of total resources controlled by the i^{th} media outlet or Web site. HHI has possible values between 0 and 10,000.

HHI and the Gini coefficient are arguably the most commonly used metrics of inequality or concen-

¹⁰Two recent cross-media studies adopt similar metrics and reach similar conclusions. Yim 2003 finds that, in traditional media, concentration increases with the number of outlets available. Comparing circulation figures of the top 100 newspapers with the number of links their Web sites receive, Hamilton suggests that the economics of producing online news may result in concentration rather than dispersion (Hamilton 2004, Ch. 7).

¹¹The Lorenz curve can be obtained by plotting the cumulative distribution function of the resource in question against the cumulative distribution of the population possessing the resource. In a population governed by perfect equality, the Lorenz curve is a perfectly straight line: 30 percent of the population owns 30 percent of the wealth, 75 percent of the population owns 75 percent of the wealth, etc.

tration in the social sciences. In this context, they are attractive in part because they differ radically in their emphases, and the relative weight they afford to observations large and small. HHI, by squaring its components, focuses on the observations with the very highest values. Smaller players receive almost no weight in calculating the final statistic, and consequently adding additional observations with a tiny share of total resources has negligible effect on the HHI. The Gini coefficient, by contrast, is a just a mean—the mean difference between the Lorenz curve and the line of perfect equality—and as such it is drawn equally from all observations in the data. Adding a large number of observations with small values raises the Gini coefficient dramatically.

For the same reasons, the Gini coefficient and the HHI scale differently with the size of the populations they measure. Imagine two villages with identical distributions of wealth, one five times the size of the other. These two towns would have identical Gini coefficients; however, the larger town would have an HHI one-fifth that of the smaller community.

5.3 Results

By both the Gini coefficient and the HHI, audiences on the Web appear at least as focused as those of traditional media. While these results need to be supplemented by additional research, they do show convincingly that the structure of online political information is hardly a radical break with the broadcast model.

Table 6 lays out these results in detail. The first column, which looks at the Gini coefficient across all media, shows that both political communities and the larger Web generate significantly higher Gini coefficients than those for radio, print, and television programming. The gap is even larger than it appears at first. Web content with a Gini coefficient of .95 actually has one-sixth the area under the Lorenz curve as magazines or newspapers, which each generate already large Gini coefficients of roughly .7.

The political communities that we explored in the previous section are each only a tiny portion of the Web. But the similar patterns emerge in global Web data, and in Weblog data on tracking unique site visitors rather than just link structure. Moreover, the Gini coefficient is the statistic of concentration which is least sensitive to changes in the size of the communities studied.

With regard to the relationship between the largest media sources and the average players, the Web

Media	Gini Coeff.	Gini, top 20	HHI
Television—Primetime Ratings	.35	.09	93
Radio—Top 50 Markets	.53	.12	19
Print—All U.S. Newspapers	.69	.25	73
Print—All U.S. Magazines	.70	.37	123
WWW—All Sites, Links	.96	.45	323
WWW—All News Sites, Traffic	n/a	.31	n/a
WWW—Weblog traffic	.89	.42	286
WWW—General politics	.93	.61	1575
WWW—Abortion	.94	.69	1754
WWW—Gun Control	.96	.64	1705
WWW—Presidency	.98	.76	3207

Table 6: This table presents data on the level of concentration across different media types. The major point to be gleaned from this table is that, both at the micro and at the macro level, the Web appears at least as concentrated as traditional media. The political communities listed here are chosen to be representative; the other eight communities crawled produced very similar levels of concentration.

is far more concentrated than any other medium. But is this apparent concentration simply because there are many more smaller players online?

The answer seems to be no. Even just among the leading outlets of every type, online audiences seem more concentrated than their offline counterparts. These findings emerge clearly in columns 2 and 3. Column 2 again presents Gini coefficients, but in this case, they are used to calculate the inequality just among the top 20 outlets in every category. In looking just at the top choices, the gini coefficients are necessarily lower, and the disparity between online and offline content narrows. By this measure online news sites are slightly less concentrated than magazines. However, by this metric every other type of online content more focused than every category of traditional media content.

Column 3, which presents the HHI for these media categories, shows much the same story. As described above, HHI amplifies the importance of the largest information sources, while ignoring the smallest outlets. All of the political communities explored produce extremely high measures of HHI. Given that these communities are by definition smaller and more homogenous than the medium as a whole, these findings are perhaps unsurprising. Yet even looking at broader online data, such as the global link structure, shows markedly higher scores for online content than for the similarly-broad content on radio and in print.

In interpreting these findings, it is important to recognize that many of these media outlets—in print, radio, and in the global Web data—are owned by larger parent companies. We are confident that

we have chosen the right level of analysis for the data at hand; after all, newspapers, magazines, and even radio stations often present identical content on their Web sites. Still, this analysis only partially address deeper questions about the “real” diversity of information sources across media types.

Yet even with this limitation in mind, the findings above are striking. The gap between the largest players and the average outlet is far more pronounced on the Web than it is in traditional media. More importantly, this disparity is not just because the Web’s lower barriers to entry let more players into the market. Even just among the most important sources of information, the Web appears to be at least as concentrated on a few winners as the media that preceded it.

6 Conclusion

The body of this paper has focused on technical subjects of a sort that scholars of politics have rarely considered. It has talked at length about the reasons that link density is an effective proxy for online audience share. It has shown that communities of Web sites on different political topics are each dominated by a small set of highly successful sites. And it has demonstrated that, according to the most widely used metrics, the patterns of concentration we see in political communities and over the Web as a whole are at least as great as those we find in radio, newspapers, and magazines.

In concluding, it is important to remind ourselves why all of this matters. Much scholarship on what the Web means for politics hinges on arguments about audience concentration. We know that the Web gives citizens millions of choices about where to go to get their political information. What we *have not* known, however, is how much this really matters—how much the Web expands the number of choices that people actually use.

The lack of definitive data has allowed scholars to make very different assumptions about the political impact of the Web. Those who have made grand claims about the Internet and politics have often argued that the Web is part of an epochal shift from broadcasting to narrowcasting. In this view, wired citizens are supposed to rely on a much broader set of sources for their political information. Our research provides no support for these utopian or dystopian visions. There is more work to be done in comparing concentration across media, and discussions to be had about the the best data sets to use and the best metrics to adopt. But on the most fundamental claims—that the Web would make the flow of political information radically less hierarchical—the verdict is clear. Yes, almost anyone

can put up a political Website. But this fact means little if most of these sites receive only a handful any visitors. Putting up a political Website is usually equivalent to hosting talk show on public access television at 3:30 in the morning.

For those who have assumed that that the Web will transform politics for good or ill, this paper thus challenges their most fundamental assumptions. But they are not the only ones for whom this research is problematic. The scale of online concentration is so profound that it forces to rethink even the skepticism that has become conventional wisdom in political science. In examining the Web, more cautious scholars have dwelt on numerous factors that might mitigate the Web's political influence. They have talked about the "digital divide," the movement of traditional news outlets and interest groups online, and the commercialization of the Web. They have suggested that Americans' disengagement from politics and their lack of political sophistication restricts the content citizens see.

Cyberskeptics have been correct in arguing that the Web is not going to move the public's attention from a few broadcast outlets to a host of small-scale Web sites, but our data suggests that they have been right for the wrong reasons. Large sites are clearly important on the Web—Yahoo dwarfs most other portal sites, Amazon.com dominates online book selling, Ebay dominates online auctions, and online news is dominated by familiar names like CNN and *The New York Times*. What scholars have *not* generally understood, though, is that these winners-take-all patterns are repeated *at every level* of the Web.

The very pervasiveness of these phenomena belies the explanations that political scientists have offered for them. We do not blame America's high rate of functional illiteracy for Amazon.com's market dominance; it thus begs credulity to think that civic shortcomings are the driving force behind concentration in the political news market. The online political communities we study are not driven by commercial pressure, and yet the winners-take-all patterns within them are stark. Nor can we blame these patterns on powerful interest groups. The increasingly-important Weblog community is noncommercial, and initially had little association with traditional political groups. And yet, Weblogs quickly came to obey the same power law distribution in both links and traffic that we see in the Web as a whole.

The clear implication is that more fundamental forces are at work—and political scientists need to understand these larger phenomena before grafting traditional models of politics onto the online

environment.

This paper is descriptive in focus, and does not seek to offer a full or complete explanation for the emergence of the online political concentration we describe. Yet part of the reason why the Web is so concentrated surely lies in the sheer size of the medium, and the inability of any citizen, no matter how sophisticated and civic-minded, to cover it all. In trying to explain why the existence of so many more sources of political information online might not change consumption patterns, skeptical political scientists have not completely ignored the cognitive limits of citizens. Still, they have talked as much about the motivations of citizens as about the unavoidable limits on their time and on their mental abilities. The suggestion has often been that the Web might fulfill much of its promise to decentralize the flow of political information if every citizen was well informed and politically engaged.

A careful reading of the evidence suggests otherwise. In most areas of political science, it is common to assume that most citizens know little about politics and that they routinely take drastic shortcuts in the processing of political information. But if strong heuristics are needed to decide between two candidates on a ballot, how much more extreme do these heuristics need to be for citizens to decide among literally millions of political Web sites? Previous scholarship on technology and politics has not emphasized enough this profound mismatch between the vastness of online political information and the quite limited time and cognitive resources citizens possess. Making this a central concern would help explain the puzzle of online concentration, and would ultimately make scholarship on technology and politics more consistent with the models found in the rest of the discipline.

Partly, then, political scientists need to develop more explicit models of how citizens respond to the astonishing overabundance of online information. At least as important, though, is the need to fundamentally reassess how the political possibilities of the Web are constrained by its architecture.

More than anything else, it was the ostensible openness of the Web that inspired political scientists to take note of it. Scholars located this openness in the Internet's most basic design decisions: the end-to-end protocol which runs the Internet allows any computer online to connect to any other; a link on an HTML page can point to anywhere on the Web. But the link structure of the Web is part of the architecture too—a critical and often overlooked part. This paper has discussed at length how link structure shapes the content that citizens see, by providing paths for surfers, by determining search engine rankings, and by ultimately funnelling traffic to a few dominant sites.

The various pieces which make up the architecture of the Web function as a whole—and that system is only as open as its most narrow chokepoint. The end-to-end nature of the Web might not limit on the political sites that citizens visit, but the link structure of the Web certainly does.

In concluding, we suggest that the online concentration we describe in this article has quite broad implications for politics. Numerous areas of political science depend critically on assumptions about the flow of political information—from interest group formation to political engagement, voting behavior to political mobilization, public opinion to partisanship, collective action problems to democratic discourse. While scholars in these areas have no intrinsic interest in the link structure of the Web, all of them have an obvious stake in the political messages that citizens see. More and more of these messages are coming from the Internet. As a recent report declares, “The Internet, a relatively minor source for campaign news in 2000, is now on par with traditional outlets such as public television broadcasts, Sunday morning news programs and the weekly news magazines” (Pew 2004). The dramatic growth of the Web as an outlet for political information will likely continue for the foreseeable future.

The Internet is becoming an increasingly substantial part of Americans’ media diet. If political scientists want to know the impact of the political messages citizens see online, they must first understand the patterns of concentration which govern almost every aspect of online life, politics very much included.

References

- Adamic, Lada A. and Bernardo A. Huberman. 2000. "The Nature of Markets on the World Wide Web." *Quarterly Journal of Economic Commerce* 1:5–12.
- Albert, A., H. Jeong and A.-L. Barabasi. 1999. "Diameter of the World Wide Web." *Nature* 401:130–131.
- Barabasi, A.-L. and R. Albert. 1999. "Emergence of scaling in random networks." *Science* 286:509–512.
- Barabasi, A.-L., R. Albert, H. Jeong and G. Bianconi. 2000. "Power-law distribution of the World Wide Web." *Science* 287:12–13.
- Barber, Benjamin. 1984. *Strong Democracy*. Berkeley, CA: University of California Press.
- Bennett, Lance. 1998. "The Unicivic Culture: Communication, Identity, and the Rise of Lifestyle Politics." *PS: Political Science and Politics* 31:740–761.
- Bimber, Bruce. 2000. "The Gender Gap on the Internet." *Social Science Quarterly* 81:868–876.
- Bimber, Bruce. 2003. *Information and American Democracy: Technology in the Evolution of Political Power*. Cambridge, UK: Cambridge University Press.
- Brin, Sergey and Lawrence Page. 1998. "The anatomy of a large-scale hypertextual Web search engine." *Computer Networks and ISDN Systems* 30:107–117.
- Burges, Christopher J. C. 1998. "A Tutorial on Support Vector Machines for Pattern Recognition." *Data Mining and Knowledge Discovery* 2:121–167.
- Cederman, Lars-Eric. 2003. "Modeling the Size of Wars: From Billiard Balls to Sand Piles." *American Political Science Review* 97:135–150.
- Cortes, C. and V. Vapnik. 1995. "Support-Vector Networks." *Machine Learning* 20:273–297.
- Dahl, Robert. 1989. *Democracy and Its Critics*. New Haven, CT: Yale University Press.
- Davis, Richard. 1998. *The Web of Politics*. London: Oxford University Press.
- Davis, Richard and Diane Owen. 1998. *New Media in American Politics*. London: Oxford University Press.
- de Sola Pool, Ithiel. 1993. *Technologies Without Boundaries: On Telecommunication in a Global Age*. Cambridge, MA: Harvard University Press.
- DiMaggio, Paul and Eszter Hargittai. 2001. "From the 'digital divide' to 'digital inequality': Studying Internet use as penetration increases." Princeton University Center for Arts and Cultural Policy Studies, Working Paper Series number 15.
- Ding, Chris, Xiaofeng He, Parry Husbands, Hongyuan Zha and Horst Simon. 2002. PageRank, HITS and a Unified Framework for Link Analysis. Technical Report No. 49372. LBNL.
- Faloutsos, Michalis, Petros Faloutsos and Christos Faloutsos. 1999. On Power-law Relationships of the Internet Topology. In *SIGCOMM*. pp. 251–262.

- Flake, Gary William and Steve Lawrence. 2002. "Efficient SVM Regression Training with SMO." *Machine Learning* 46:271–290.
- Gini, Corrado. 1921. "Measurement of Inequality of Incomes." *The Economic Journal* 31:124–126.
- Graber, Doris A. 1996. "The 'New' Media and Politics: What Does the Future Hold?" *PS: Political Science and Politics* 29:33–36.
- Hamilton, James T. 2004. *All the News That's Fit to Sell: How the Market Transforms Information into News*. Princeton, NJ: Princeton University Press.
- Hargittai, Eszter. 2003. "How Wide A Web?: Inequalities in Accessing Information Online." Doctoral Dissertation. Department of Sociology: Princeton University, Princeton, NJ.
- Huberman, Bernardo A. 2001. *Laws of the Web*. Cambridge, MA: MIT Press.
- Huberman, Bernardo A., Peter L. T. Pirolli, James E. Pitkow and Rajan M. Lukose. 1998. "Strong Regularities in World Wide Web Surfing." *Science* 280:95–97.
- Joachims, Thorsten. 1998. Text categorization with support vector machines: learning with many relevant features. In *Proceedings of ECML-98, 10th European Conference on Machine Learning*, ed. Claire Nédellec and Céline Rouveirol. Chemnitz, DE: Springer Verlag, Heidelberg, DE pp. 137–142.
- Karmark, Elaine Ciulla and Joseph S. Nye, eds. 2002. *Governance.com: Democracy in the Information Age*. Washington D.C.: Brookings.
- Kleinberg, Jon M. 1999. "Authoritative sources in a hyperlinked environment." *Journal of the ACM* 46:604–632.
- Krugman, Paul. 1994. "Complex Landscapes in Economic Geography." *American Economic Review* 84:412–416.
- Kumar, Ravi, Prabhakar Raghavan, Sridhar Rajagopalan and Andrew Tomkins. 1999. "Trawling the Web for emerging cyber-communities." *Computer Networks (Amsterdam, Netherlands: 1999)* 31:1481–1493.
- Lawrence, Steve and C. Lee Giles. 1998. "Searching the World Wide Web." *Science* 280:98–100.
- LeCun, Y., L. Jackel, L. Bottou, A. Brunot, C. Cortes, J. Denker, H. Drucker, I. Guyon, U. Muller, E. Sackinger, P. Simard and V. Vapnik. 1995. "Comparison of learning algorithms for handwritten digit recognition." *International Conference on Artificial Neural Networks*.
- Liljeros, Fredrik, Christofer R. Edling, Luis A. Nunes Amaral, H. Eugene Stanley and Yvonne Aberg. 2001. "The Web of Human Sexual Contacts." *Nature* 411:907–908.
- Lupia, Arthur and Gisela Sin. 2003. "Which Public Goods Are Endangered?: How Evolving Communications Technologies affect *The Logic of Collective Action*." *Public Choice* 117:315–331.
- Manjoo, Farhad. 2003. "Can the Web Beat Big Media?" *Salon*. May 21.
- Marendy, Peter. 2001. A Review of World Wide Web searching techniques, focusing on HITS and related algorithms that utilise the link topology of the World Wide Web to provide the basis for a structure based search technology. Technical Report. James Cook University. North Queensland, Australia.

- Nielsen-Netratings. 2004. Nielsen NetRatings Search Engine Rankings. Technical Report. Search Engine Watch.
URL: <http://searchenginewatch.com/reports/article.php/2156431>
- Norris, Pippa. 2001. *Digital Divide: Civic Engagement, Information Poverty and the Internet in Democratic Societies*. New York: Cambridge University Press.
- NTIA. 2000. Falling Through the Net: Toward Digital Inclusion. Technical Report. National Telecommunications and Information Administration.
- NTIA. 2002. A Nation Online: How Americans Are Expanding Their Use of the Internet. Technical Report. National Telecommunications and Information Administration.
- Osuna, E., R. Freund and F. Girosi. 1997. “Improved training algorithm for support vector machines.” Technical Report. NNSP’97.
- Pandurangan, Gopal, Prabhakara Raghavan and Eli Upfal. 2002. Using PageRank to Characterize Web Structure. In *8th Annual International Computing and Combinatorics Conference (COCOON)*.
- Pareto, Vilfredo. 1897. *Cours d’Economie Politique*. Vol. 2.
- Pennock, David M., Gary W. Flake, Steve Lawrence, Eric J. Glover and C. Lee Giles. 2002. “Winners Don’t Take All: Characterizing the Competition for Links on the Web.” *Proceedings of the National Academy of Sciences* 99:5207–5211.
- Pew. 2004. Cable and Internet Loom Large in Fragmented Political News Universe. Technical Report. The Pew Research Center for People and the Press.
- Platt, J. 1998. Sequential minimal optimization: A fast algorithm for training support vector machines. Technical Report No. 98–14. Microsoft Research, Redmond, Washington, April 1998.
URL: <http://www.research.microsoft.com/jplatt/smo.html>
- Shapiro, Andrew L. 1999. *The Control Revolution*. New York, NY: Public Affairs.
- Shirky, Clay. 2004. “Inequality in the Weblog World.” Seminar Presentation, Berkman Center for Internet and Society.
- Sunnstein, Cass. 2001. *Republic.com*. Princeton, N.J.: Princeton University Press.
- Vapnik, Vladimir. 1995. *The Nature of Statistical Learning Theory*. New York: Springer.
- Wilhelm, Anthony G. 2000. *Democracy in the Digital Age: Challenges to Political Life in Cyberspace*. London, UK: Routledge.
- Yim, Jungsu. 2003. “Audience Concentration in the Media: Cross-media Comparisons and the Introduction of the Uncertainty Measure.” *Communication Monographs* 70:114–128.

A Appendix on Methodology

A.1 Support Vector Machine Classifiers

In order to implement the research design described in this article, it is essential to have a reliable method of automatically classifying Web pages. We solve this problem with the use of a support vector machine classifier.

Support vector machines are a learning theory method introduced by Vapnick et al. (Cortes and Vapnik 1995, Vapnik 1995). SVM techniques have received a good deal of attention from computer scientists and learning theorists in recent years,¹² and have found uses in a wide variety of applications—from face detection (Osuna, Freund and Girosi 1997) to handwritten character recognition (LeCun et al. 1995). But support vector machines are particularly effective in classifying content based on text features—an area where SVM methods show substantial performance improvements over the previous state of the art, while at the same time proving to be more robust (Joachims 1998).

Mathematically, support vector machines are a technique for drawing decision boundaries in high-dimensional spaces. In low numbers of dimensions, and with a straight line as the decision boundary, it is relatively simple to visualize and understand how SVM’s operate. In Figure 5, for example, one can see a plot containing two different types of points. The circles are clustered in the lower left-hand corner of the plot, the squares in the upper right corner. These two groups of points are the “training set”—the initial set of points that teach the SVM where to draw the boundary separating the two groups. Now consider only the points closest to the boundary line. Each of these points is a *support vector*.

The decision boundary is drawn in an attempt to maximize the distance between the support vectors. In this example, this maximization defines the slope of the straight line separating the two groups of points, in much the same way as minimizing the sum of squared errors defines the slope in an OLS regression. Unlike regression analysis, however, SVM’s deliberately avoid using all of the information available. The number of support vectors is generally quite small; and while the problem is still computationally intense, it is markedly less so than most feasible alternatives. Once the boundary line is drawn, newly encountered points can be classified by their position in this space. In our example, the SVM would assume that any new point above the line was a square, and any point below the line

¹²For an accessible and widely-cited introduction to support vector machines, see Burges 1998.

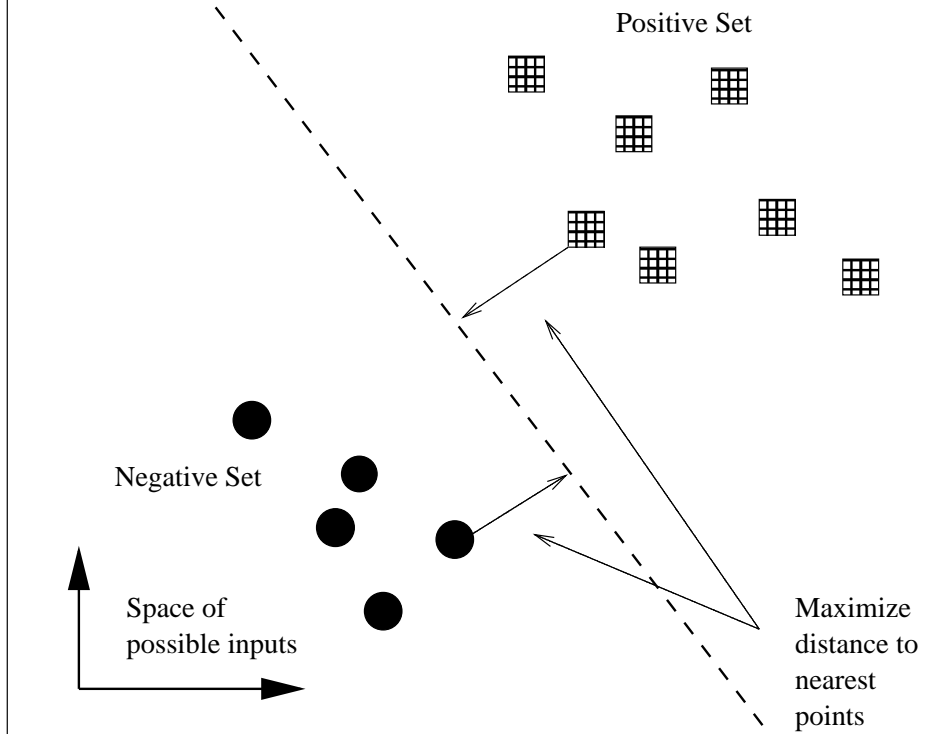


Figure 3: This figure shows a simple linear support vector machine. The boundary decision line is drawn to maximize the distance between itself and the *support vectors*, the points closest to the line. This example owes much to the explication of Platt 1998.

was a circle.

In our analysis, SVMs work by converting the HTML document representing a particularly Web page into a single point in a high-dimensional space; the decision boundary is represented by a hyper-plane cutting through this space. The HTML document is broken up into a series of *features*, which are either words or word pairs. Each feature is a dimension. The document’s value on this dimension is 1 if the feature—for example, the phrase “United States”—occurs in the given page; otherwise the value is zero. One of the primary advantages of SVMs is that the difficulty of learning depends on the complexity of drawing the appropriate margin, and is only indirectly related to the dimensionality of the feature space. For the purposes of this paper, we implement sequential minimal optimization (SMO) in order to train our support vector machine (Platt 1998, Flake and Lawrence 2002).

A.2 Surfer Behavior and Crawl Depth

In this study we examine all pages that are three clicks or less away from our seed sets. It is worth a brief detour to explain why travelling only three links away from the seed set should capture the large majority of relevant political Web sites.

The diameter of the Web is small: two randomly chosen Web sites are, on average only 19 hyperlinks apart (Albert, Jeong and Barabasi 1999). By traveling three links away from our seed set, our study

examines graphs with a diameter of 6—three links in any direction. One consequence of this property, however, is that crawling more than a few links away from the original seed set requires crawling a large fraction of the World Wide Web. In this case, increasing the depth of the crawl by 1 increases the number of sites that must be downloaded, stored, and analyzed by a factor of 20.

Research on the behavior of Web surfers gives us strong reason to believe that increasing the depth of the crawl would be of limited benefit. Huberman et al. show that the number of links that a user will follow away from a starting Web site can be modeled extraordinarily well by an inverse Gaussian distribution. The probability that any path on the Web will exceed depth L is governed by the following equation:

$$P(L) = \sqrt{\frac{\gamma}{2\pi L^3}} \exp\left[-\frac{\gamma(L - \mu)^2}{2\mu^2 L}\right]$$

Data taken from the unrestricted behavior of AOL users produces estimates of γ and μ of 6.24 and 2.98, respectively. While most surfing paths on the Web are only a few clicks deep, the heavy tails of the Gaussian distribution mean that even a path that contains a dozen or more clicks contains a non-trivial portion of the probability mass.

This research suggests that the moderately deep crawl we perform should capture the large majority of surfing behavior away from the seed sites. If Huberman's numbers hold, roughly 80% of searches will terminate before exceeding the depth of the crawl we perform. The benefits of a deeper crawl appear to be modest. Increasing the depth one level would expand the portion of search behavior covered by only 5–10%, while it would increase the difficulty of analysis by a factor of 20. To provide a sense of perspective, increasing the depth of the crawl by one would have required us to download and analyze 4.5 million Web sites for *each* of the 12 crawls. This would have meant crawling roughly 54 million pages total, and would ultimately have taken up more than 5 terabytes of disk storage.