"Googlearchy": How a Few Heavily-Linked Sites Dominate Politics on the Web*

Matthew Hindman† Kostas Tsioutsiouliklis‡ Judy A. Johnson§

July 28, 2003

Abstract

Claims about the Web and politics have commonly confounded two different things: retrievability and visibility, the large universe of pages that could theoretically be accessed versus those that citizens are most likely to encounter. While the governing assumption of much previous work has been that retrievability would translate inexorably into visibility, we cast doubt on that claim. Drawing on a large literature in computer science that ties a site's visibility to the number of inbound hyperlinks it receives, this paper proposes a new methodology for measuring the link structure surrounding political Web sites. Our technique involves iterative, extremely large-scale crawls away from political sites easily accessible through popular online search tools, and it uses sophisticated automated methods to categorize site content. In every community we examine, we find that a small handful of Web sites dominate. Online political communities on the Web thus seem to function as "winners take all" networks, a fact that would seem to have widespread implications for politics in the digital age.

*An early version of this paper was presented April 4, 2003, at the Annual Meeting of the Midwest Political Science Association, Chicago, IL. Adam Simon, Arthur Lupia, Adam Berinsky, and other panel participants provided numerous helpful comments. Jennifer Hochschild and Paul DiMaggio deserve thanks for their important and timely insights on the piece; Kenn Cukier's thoughts, editting, and unflagging support were invaluable. Lastly, we gratefully acknowledge Lada Adamic's contribution in providing us with data on the relationship between inbound links and site traffic.

†Ph.D. candidate, Department of Politics, Princeton University. (Doctoral Fellow, National Center for Digital Government, Kennedy School of Government, 79 John F. Kennedy St., Cambridge MA 02138; http://hindman.cc/, mhindman@princeton.edu). Correspondence regarding this piece should be addressed to Matthew Hindman; email correspondence strongly preferred.

‡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

On June 2, 2003, the Federal Communications Commission voted to loosen restrictions on media ownership which dated back to the early 1940s. The stated rationale for this important change was that cable and satellite television—and, even more importantly, the advent of the Internet—now ensure the diversity of political information sources better than any amount of government regulation. As commission chairman Michael K. Powell declared with emphasis on the Internet, "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)).

The FCC's decision demonstrates the growing impact that our ideas of the Internet have on both the theory and the practice of politics. Powell's claim that technology is now "a defining tenet around which to organize our thinking, our industrial activity, and our conception of the public interest" (Manjoo (2003)) may seem like a remnant of the dot-com boom's cyber utopianism. But for scholars of politics, it is worth noting that technological optimism in the discipline has a pedigree that predates the modern Web.[1] In *Strong Democracy* (1984), Benjamin Barber looked to new information technologies to enable much higher levels of direct citizen participation. Robert Dahl, arguing that information disparities are more dangerous than economic ones, wrote hopefully in *Democracy and Its Critics* (1989) that new information technologies might ameliorate political inequalities by making political information more widely available. Amitai Etzioni, in *The Spirit of Community* (1993), argued that information technology would enable citizens to build stronger communities and a more politically egalitarian future.

The Internet is no longer an object of far-fetched speculation, but an indispensable part of the nation's economic and social life. More than 54% of the U.S. population now uses the Web; all but six states have a majority of their citizens online (NTIA (2002)). Still, many scholars have felt not just that the medium has failed to live up to initial hopes, but that its social impact is now a source of deep concern. Cass Sunnstein has argued that the Web will lead to increased isolation, balkanization, and the loss of a common

[1] This characterization of the prehistory of political science optimism about information technology is indebted to Bruce Bimber's discussion in *Information and American Democracy* (Bimber (2003)).

political discourse (Sunnstein (2001)). Benjamin Barber, comparing the actuality of the new medium with his early hopes, has ended up voicing much the same worry (Barber (1998)). Joseph Nye suggests, or at least takes seriously, the idea that "the demise of broadcasting and the rise of narrowcasting may fragment the sense of community and legitimacy that underpins central governments" (Karmark and Nye (2002), p. 10).

Though empirical examinations of the Web have generally been more reserved than theoretical treatments, they have also offered ample criticism of the notion that new information technologies would foster unfettered political equality. Much effort has been expended in chronicling social barriers which limit equality on the Internet. Research on the "digital divide" has examined the extent to which Internet adoption follows traditional social cleavages such as income, education, gender, age, and race (NTIA (2000); Bimber (2000); Hoffman and Novak (2000)). Other scholars have examined profound gaps in the skills users need to use the medium effectively (DiMaggio and Hargittai (2001); Hargittai (2003)). In response to claims that the Internet would lead to dramatically increased openness, some have argued that the movement of traditional media firms online perpetuates previous patterns of influence (Davis (1998); Davis and Owen (1998)).

Scholarly perspectives on the Internet thus seem to have covered quite a bit of ground. Even so, one feature of these debates is curious. Early optimism about information technologies and contemporary dystopian visions disagree on most points, but both take as given that a greater diversity of political information is an inevitable consequence of the Internet. Though the recent FCC decision occasioned much public dissent, almost no one questioned Chairman Powell's claim that, for those able to access it, the Internet is an astonishing well of diversity. Even the most sober social science inquiries have often valorized the openness of the medium's architecture. By focusing on social barriers to access and effective Web use, the implication is often that the promise of the Web would be fulfilled if only offline social inequality could be overcome. Even when this is not the case, skeptics have lacked systematic data with which to challenge the perceived openness of the medium. Indeed, it is usually the ostensibly egalitarian nature of the Internet that inspires scholarly interest in the first place. In Lupia and Sin's recent study of the Web's impact on collective action problems, for example, it is the radical accessibility of online content which motivates their discussion:

> 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 (Forthcoming)).

These sort of statements about the nature of the Internet may seem commonsensical, even blindingly obvious. In truth, they are deeply problematic. Discussions about the Web often tout how it increases the "availability" or "accessibility" of information. But in doing so, they usually confound two very different things: what we might term *retrievability* and *visibility*. Scholars, policymakers and popular commentators rightly point out that the Net provides equality of retrievability; the Web's "end-to-end" architecture guarantees that any page online can be accessed by any other computer on the Internet. What is far more important for politics, however, is the visibility of different Web sites—what pages citizens are actually likely to encounter, either when deliberately searching for political information or in the course of their ordinary online activities. While retrievability online is absolute (a page can either be accessed or it cannot), visibility is relative. A personal home page may be as retrievable as `yahoo.com`, but is certainly not as visible.

Making the distinction between retrievability and visibility highlights an enormous gap between the claims made about the Internet's effect on politics, and much of the evidence marshalled to support them. Statements that the Web will transform mass politics for good or ill, or that it will be a boon for information diversity, cannot be sustained just on the openness of the Web's technical architecture. They assume both greater openness in the posting of political information, and a corresponding decrease in the concentration of viewership. In other words, statements about the Web rest on the implicit belief that retrievability will translate—imperfectly perhaps, but inevitably—into visibility. But political scientists have yet to examine if, or to what extent, this belief is true.

Some insight into this question, however, can be gleaned from research in other disciplines. Computer scientists and applied physicists have been much quicker to examine patterns of visibility and traffic on the Web than their peers in the social sciences. What they have found may cause us to doubt common assumptions. There is now a large and

mature body of scholarship demonstrating that, for the Web as a whole, both site traffic and the number of inbound hyperlinks that a site receives is power law distributed (Adamic and Huberman (2000); Broder et al. (2000); Barabasi and Albert (1999); Kumar et al. (1999)). In practice, this means that a tiny fraction of sites dominate the information environment.

Political science has generally overlooked the knowledge that computer scientists and applied physicists have gathered about the structure of the Web. Discussions about power laws and inlink distributions may seem arcane, and divorced from the concerns of political scientists. Nothing could be further from the truth. Many people know that Google, the Web's dominant search engine, ranks pages based on their link structure (Brin and Page (1998); Pandurangan, Raghavan and Upfal (2002)). Less well known, however, is that most other modern search engine algorithms—even those radically different from Google's PageRank—end up returning heavily-linked sites first (Ding et al. (2002)). Moreover, we present user data showing that the number of inbound links that a site receives is a good predictor of its traffic—because there are more online paths to heavily-linked sites, because high-quality sites tend to attract more links, and because much Web traffic is generated by search engines. If we want to understand the visibility of political Web sites—and thus the ability of the Web to reorder the economy of political information—we need to examine the hyperlinks that connect these sites to the rest of the Web.

Given that link structure tells us a great deal about which pages are easily found and likely to be seen, this paper proposes to measure link structure on a massive scale. We present a new methodology to analyze hyperlinks surrounding political Web sites, one which takes advantage of both recent advances in theoretical computer science and powerful hardware and software resources provided by a large corporate research laboratory. We use the two most popular online search tools, Yahoo and Google, to create "seed sets" using their most highly-ranked results in six diverse categories of political Web sites. We then crawl outward from each of these seed sites, using support vector machines to classify newly encountered pages as relevant to a given category of political content. Our methods scale well to deal with the vast size of the Web; the twelve crawls we perform require us to download, on average, a quarter of a million Web pages each, or almost 3 million pages total. Our techniques produce estimates of the total amount of political content

easily reachable from these important search tools, and they highlight the pages ordinary citizens are most likely to see.

At stake in this analysis is a host of claims about the Internet's relationship with politics. The hope, and sometimes the fear, has been that the Web would promote equality in both the dissemination and reception of information—allowing ordinary citizens to both post their views prominently, and to consume diverse sources of political information posted by others. Though no one expected that every page on the Web would receive an exactly equal share of attention, many have assumed that the Web would be dramatically more egalitarian in this regard than traditional media. Our empirical results, however, suggest enormous disparities in the number of links pointing to political sites in a given category. In each of the highly diverse political communities we study, a small number of heavily-linked sites receive more links than the rest of the sites combined—effectively dominating the the community they are a part of.

Whether this result is surprising depends on one's prior expectations—and in this area, expectations seem to be very diverse indeed. Researchers who discovered global power law structures on the Internet initially found them quite surprising (Barabasi (March 3, 2003)), and it has been an open question whether most online communities are governed by power laws at the micro level. Our finding run counter to previous research, which found that a number of prominent subcommunities on the Web—universities, newspapers, publicly listed companies—have link structures that follow very different, log-normal distributions (Pennock et al. (2002)). At the same time, power law structures have been previously observed in a wide array of other social phenomena—from the distribution of income (Pareto (1897)) to the frequency of words (Zipf (1949)), from the intensity of wars (Richardson (1948); Cederman (2003)) to the distribution of sexual contacts (Liljeros et al. (2001)).

We introduce a new term to describe the organizational structure we find: "googlearchy"—the rule of the most heavily linked. We ultimately conclude that the structure of the Web funnels users to only a few heavily-linked sites in each political category. Claims about greater information diversity online cannot be evaluated without more thorough comparative data on offline media. What our findings do make clear, however, is that the links structure of the Web puts significant limits on the medium's ability to democratize political information.

2 What Link Structure Can Tell Political Scientists

The Web is big—*very* big. The indexable Web now includes billions of pages, more than a single researcher could explore in many lifetimes. 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. Researchers have tried a number of different tactics to sidestep this problem. They have drawn deductive conclusions about the Web's social impacts based on the openness of its architecture. They have examined important case studies, performed small-scale content analyses, watched Web users in laboratory environments, catalogued the ways that specific groups have used the Internet to organize themselves. And of course they have conducted surveys of Internet users, examining their demographic characteristics and broad patterns of usage.

All of these methods have produced important insights about the Web's relationship to politics. Yet all have limitations; the Web is so large that even surveys with several thousand respondents tell us little about any single category of political information. The 2000 General Social Survey—to date the most comprehensive cross-sectional study of Web usage—is a case in point. Of 2817 total respondents, standard research design limitations reduce to 200 the number of respondents who report visiting a political Web site in the last 30 days. Even if more specific questions about the sites visited had been asked, the most popular subcategories of political sites would be left with only a handful of visitors, a tiny fraction of the data needed to make statistically valid inferences.

In this paper we present a new, different, and complementary approach: a methodology for analyzing the structure of the hyperlinks leading to and from political Web sites. This methodology, while innovative, is grounded theoretically by an extensive literature in computer science and applied physics. Though comprehensive analysis of link structure requires software and hardware resources not commonly available to social scientists, it is nonetheless much easier and cheaper to perform than alternative approaches such as large-scale surveys. All of the data it draws on is, by definition, publicly available. It does not require the use of sampling; in fact, it can catalog *all* Web pages easily reachable using certain online tools given specified constraints, even when the number of pages to

be explored numbers in the hundreds of thousands. And it allows us to make inferences about the relative visibility of Web sites even within topical communities that would never show up in a cross-sectional survey—critical if we are to take seriously the notion that the Web is a "pointcasting" or "narrowcasting" medium. Link data has limitations of its own, and there are important questions which it cannot answer. But if we want to understand the Web's ability to make political information less concentrated, the link structure of the Web represents an extraordinarily rich, and heretofore untapped, resource.

To place this approach in context, it is worth reviewing what we already know about the distribution of hyperlinks online. The structure of the Web has been a remarkably fertile area of scholarship in recent years, and researchers have discovered that links between sites obey very strong statistical regularities. More specifically, when looked at over the entire Web, the distribution of both inbound and outbound hyperlinks follows a power law or scale-free distribution over many orders of magnitude (Barabasi and Albert (1999); Kumar et al. (1999)). Put more precisely, the probability that a randomly selected Web page has K links is proportional to $K^{-\alpha}$ for large K.

To get a sense for what such a distribution looks like in practice, imagine a hypothetical community where wealth is power law distributed. At one end of the spectrum, there is one billionaire, ten individuals worth at least 100 million dollars, a hundred people worth 10 million dollars, and a thousand people worth at least a million. At the opposite end of the spectrum, 1,000,000 people have a net worth of $1,000. In this hypothetical community, wealth is distributed according to the function $K^{-\alpha}$, where $\alpha = 1$.

Empirical studies that have tried to measure the dimensions of this effect on the Web have shown that $\alpha \approx 2.1$ 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)).[2] Intuitively, this finding means that links on the Web are distributed in an extremely inegalitarian fashion. A few popular sites (such as Yahoo or AOL) receive a huge portion of the total links; less successful sites (such as most personal Web pages) receive hardly any links at all.

It may not be immediately obvious that these findings by computer scientists have

[2]Barabasi et al. and Kumar et al. seem to disagree on the value of α for outgoing hyperlinks; Barabasi et al. propose a value of $\alpha = 2.4$.

anything to do with the concerns of political scientists. But consider again what it means to have a piece of political information be visible. Highly visible political information is, quite simply, information that is easily found. On the Web, new content can be found in 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.

2.1 The Relation Between Inbound Links and Web Traffic

We know that both Web traffic and Web links are power law distributed (Huberman et al. (1998); Adamic and Huberman (2000)). But how close is the relationship between link structure and site visits? We present below a new analysis of an older data set showing that the relationship between the two is strong.[3]

The data set consists of only two variables: the number of visits that a Web site receives, and the number of inbound links leading to that site. 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; the crawls on which the link data is based were performed in late 1997.

Given the simplicity of the data set, analysis is straightforward. The number of hyperlinks pointed to a site and the number of visits it receives are highly correlated, with a correlation coefficient of .698. The number of hyperlinks pointing to a site does seem to be a good predictor of its traffic.

This data should be interpreted with caution, for several reasons. First of all, it is from an earlier period in the Web's history, and the medium has changed and expanded a great deal in the meantime. However, all indications are that, if anything, the connection between traffic and link structure has grown much stronger in the intervening years. At the time this data was collected, the most popular search engines, such as Alta Vista, paid little attention to the link structure surrounding Web pages. Since contemporary search engines now focus heavily on link structure, and since search engines drive a large portion of Web traffic, it would be very surprising indeed if the link between traffic and

[3]We thank Lada Adamic for providing us with the data analyzed below.

link structure had waned instead of waxed.

Second, a few prominent sites in the data are ad sites, which receive numerous inbound links but, unsurprisingly, a relatively small share of traffic. Thus, for non-ad sites, the correlation between links and traffic is likely higher than that reported above. However, because the terms under which we have been provided the data do not permit the individual identification of Web sites, advertising sites cannot be removed from the analysis.

Third, most of the variance in both links and site visits comes from a small number of sites with large values. The correlation figure can thus change markedly with the omission or inclusion of one of the most successful sites. It is an essential feature of power law distributions, however, that they have most of their variance in a small number of heavily-leveraged observations.

2.2 A Formal Model of Web Surfing

It thus seems clear that the amount of Web traffic on a site is highly correlated with the number of inbound hyperlinks which point to that site from the rest of the Web. Increasingly, though perhaps not in the early data above, the relationship between site visits and links structure is reinforced by modern search engines. As will be explained at greater length below, most search engines in practice end up ranking sites based on the number of inlinks these sites receive—even when the underlying ranking algorithm is far more complex. Since much Web traffic originates from these search engines, traffic follows links.

Still, as was clear even early on in the Web's history, much of the association between inbound links and traffic is simple: hyperlinks exist to be be followed. The more hyperlinks there are to a given site, the more chances users on connecting sites have to follow them, and the more traffic the site ultimately receives. And once this traffic is generated, much of it is passed along to "downstream" sites. A single link from, say, the popular online journal `Slashdot.org` can generate more traffic than links from hundreds of less prominent sites.

Thus, even without relying on assumptions about search engines generating traffic, or about traffic to valuable sites generating additional links through a recursive process, a site's traffic may be a direct function of its link structure. One way of understanding the link between visibility and traffic is to formalize the intuitions above. And so we present

a highly simplified model of Web surfing behavior.[4]

Let the "Web" be represented as a graph of N interconnected nodes $S_1 \ldots S_N$, each of which represents a Web site. Each site S_i contains a set of directed edges—"hyperlinks"—which connect it to other sites on the Web. Now imagine a surfing agent, A. The agent A begins anywhere on the graph, say at site S_q. At each turn T, the agent follows one of the outgoing hyperlinks to another site in S. At each new site, the process is repeated, generating a random walk over the Web. It is quite likely that in our simulated Web, as in the real one, there are pockets of self-referential links from which no exit exists. To accommodate this, we make the following refinement. We add a fixed probability that A, instead of exiting the site via the outgoing hyperlinks, will instead be "teleported" at random to another site on the Web.

Our ultimate concern is the number of visits that Web page S_i will receive; this is denoted by the quantity V_i. The decay factor—the odds that our surfer will continue the random walk, instead of being automatically teleported to another node in our Web—is given by p. While p can be set at any value between 0 and 1, in this case we set p to .85. Let In_i denote the set of inbound hyperlinks which connect to site S_i, and $d_{out(i)}$ denote the number of outbound hyperlinks for site S_i. The proportion of visits that any one site should receive is thus given by the following equation:

$$V_i \propto \frac{(1-p)}{N} + p * \sum_{j \in In_i} \frac{V_j}{d_{out(j)}}$$

The end result is a model where the expected number of hits on a given Web page is a linear function of only two components. First, each page has a small, fixed chance of being teleported to. The second, and far larger, component is a direct function of link structure. It suggests a recursive function in which sites that are heavily linked to, by other sites that are also heavily linked to, receive more than their share of Internet traffic.

To repeat: visible political information online is information that can be reached easily through Web surfing activities, or information that can be found with little effort using online search tools. The model above, though quite simple, does offer insight into the former activity. It helps explain why the link structure of a site—and particularly the

[4]This model is derived from one presented by Brin and Page (1998). The full reason for reproducing it here will be presented in the next section. See also Pandurangan, Raghavan and Upfal (2002).

number and popularity of the inlinks it possesses—plays an enormous role in what online sources of political information receive attention.

2.3 Tools For Searching

All of this brings us to the second category of highly visible political information: information that can be easily retrieved using popular search tools. First, though, we must offer a confession. The formal model presented above is a plausible simplification of much Web activity, and explains many observed features of Web usage. It is also, however, the classic exposition of the PageRank algorithm, the central feature of the Google search engine (Brin and Page (1998); Pandurangan, Raghavan and Upfal (2002)).

Why do we explain PageRank first in a different context? Because we wanted to emphasize a critical point: *surfing behavior, search engine results, or any combination of the two all produce similar biases in the attention given to various Web sites.* As Google itself is based on a formal model of Web surfing, we should expect that both surfing away from known sites and the use of search tools should privilege the same set of Web pages. Sites which are heavily linked to by other prominent sites become prominent themselves; other sites are likely to be ignored. The tendency of surfers to "satisfice"—to stop after the first site that contains the sort of content sought, rather than looking for the "best" result among hundreds of relevant sites returned—makes this "winner take all" phenomenon even stronger.

Overall, 83% of the searches performed on the Web use the Google engine (Nielsen-Netratings (2003a)), and the dominance of Google makes it an attractive target for criticism.[5] One might think that a greater diversity of search engines would help ensure diversity in the content seen. But in truth, the popularity contest dynamics associated with Google and PageRank are difficult to avoid. The HITS algorithm is perhaps the most plausible alternative to PageRank, and uses the mutually reinforcing structure of "hubs" and "authorities" to determine 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

[5]Note that the 83% figure includes search portals, such as AOL and Yahoo, which use Google's technology to power their own searches. Yahoo has recently stated its intention to abandon this practice, replacing Google's engine with a proprietary alternative powered by its newly acquired subsidiaries Inktomi and Overture. Yahoo is currently used for slightly less than 30% of Web searches. For a more thorough breakdown of search engine market share, see Nielsen-Netratings (2003a).

the search engine spectrum" from PageRank, it produces nearly identical results. Indeed, both engines—and any likely competitors—produce results that are hardly different than just ranking sites by their inlink degree (Ding et al. (2002)). Google's use of PageRank to produce a weighted centrality measure is thus not as consequential as many think. No matter what search engine is used, the small number of sites with large numbers of inbound hyperlinks are returned first.

2.4 Open Architecture, Unequal Results?

We therefore know both that the number of inbound hyperlinks attached to a Web site is a central determinant of the visibility of the information it contains, and that the distribution of these inbound links over the whole Web pushes users toward small numbers of hyper-successful sites. By way of analogy, social scientists would never assume that equality of opportunity in the economic sphere would result in an equal distribution of wealth. But some observers have made a similar sort of error with regard to the Web—they have taken the open architecture of the Internet as a promise that the outcome would be similarly egalitarian. Ironically, for the Web as a whole, the fact that anyone can place information online creates problems of scale that only a few of the most successful sites may be able to overcome.

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, within all business homepages, all university homepages, and all newspaper homepages—researchers have found that the distribution of hyperlinks obeys a unimodal, 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 to a large extent 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.

3 Methodology: Gathering the Data

3.1 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.

Our technique for mapping the network of easily accessible political information can be broken down into a series of simple steps. First, we create lists of highly-ranked political Web sites in a variety of different categories; these become our "seed sites." With these seed sites in hand, we use Web robots—automated programs which act like Web users—to crawl the Web. These robots (or "spiders" or "crawlers") start at each of the seed sites, and then follow all of a site's outgoing hyperlinks, downloading all of the pages accessible from a given seed site. These downloaded pages are then classified as either "positive" (similar to pages in the seed set) or "negative" (more similar to a reference collection of random content).

This process can be iterated, as all of the links on the newfound sites are followed in turn. The "depth" of the crawl reflects the number of iterations of this process, and thus how many hyperlinks away a site can be from one of the original seed sites.

3.2 Support Vector Machine Classifier

An obvious and crucial prerequisite for successfully implementing the research design above is a reliable method of classifying newly-encountered sites as relevant to a given category of political content. Clearly, the Web is very, very large. Even aside from questions about subjectivity, it is not feasible to use human coding to classify millions of Web pages. We solve this problem with the use of a support vector machine classifier.

Details on the operation of the classifier can be found in the appendix. But it should be noted, generally, that the SVM classifier offers two advantages for our purposes. First, it can be trained with relatively little human intervention. After being provided with both

a positive set (in this case the seed set) and a negative set (a collection of random Web pages), the SVM inductively learns to differentiate between relevant and irrelevant pages.

Second, the support vector machine classifier produces highly reliable classification of Web pages. Most importantly, it produces very few false positives. Randomized human coding of SVM-classified sites shows that approximately 98% of sites in the positive set are correctly classified. Human coding also suggests that only a tiny portion of sites in the negative set are incorrectly categorized. The third category, which the SVM classifier categorizes as "unsure," seems the only potential source of problems. Human coding suggests that most sites about which the SVM classifier is unsure should be placed in the positive set. However, including these sites in our analysis would not affect the reported results.

3.3 Choosing Seed Sites

For both the Web crawling and the automated classification, much depends on the seed sites chosen. The preceding discussion on the structure of the Web gives some insight into the reasons this proves an easy problem to solve. A small handful of sites handle the bulk of search behavior: Yahoo, MSN, AOL, and Google. In this paper, we look at both human-categorized Web directories and at results returned from search engines. Yahoo's human categorized directories are the most popular content of that type, and so its categorized content is used for half of the crawls (Nielsen-NetRatings (2003b)). The dominance of Google makes it the obvious source to use for search engine queries (Nielsen-Netratings (2003a)).

We chose six categories of political content to examine for the purposes of the paper, with parallel seed sets taken from both Yahoo and Google. First, we look at sites devoted to the most general concerns of U.S. politics, looking at Yahoo's "U.S. Politics" category and Google's top results for the query "politics." Second, we look at results for broad searches about the federal government. One pair of seed sets focuses on the current President; another pair contains sites related to the U.S. Congress. Third, three pairs of seed sets contain content about longstanding, controversial political issues: abortion, gun control, and capital punishment.

Seed sets in each category are limited to 200 sites, both for Google and the Yahoo

directory. While this limit is introduced largely to provide a sense of scale across different searches, it also results from practical considerations. Google results in many of these categories degrade noticeably in quality after the first 200 results, and may wander away from the community of sites being investigated. Yahoo categories focusing on political issues are much smaller than those focusing, for example, on the U.S. Congress; exceeding 200 seed sites in many cases would have required sites to be gleaned from other sources.[6]

3.4 Surfer Behavior and Crawl Depth

Critical to this methodology is ensuring that our approach will catalog most of the political content that users of these search tools are likely to see. For the purposes of this study we thus crawl each seed site to a depth of 4, three clicks away from our seed set. It is worth a brief detour to explain why travelling three clicks away from the seed set should capture the large majority of relevant political Web sites.

The crawl depth in this study is chosen for both theoretical and practical reasons. First of all, it is well-known that the Web obeys "small world" properties, and that 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—infeasible without any extraordinary amount of infrastructure. 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. Even at a depth of 4, each search requires us to download and classify roughly a quarter of a million pages.

Aside from the hardware limitations, however, 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

[6]Yahoo results are categorized in descending topical trees, with the most popular and general sites reported first, and more specific and less popular sites relegated to subcategories. Yahoo categories on a given topic were crawled to the first depth that exceeded 200 results, and then the most recent level crawled was cut to the number of sites required to fill out the data set. For example, if a depth 2 crawl returned 150 results, and a depth 3 crawl returned an additional 100, every other site at depth 3 was included in the data set.

starting Web site can be modeled extraordinarily well by an inverse Gaussian distribution. Indeed, the probability that any path on the Web will exceed depth L is governed by the following equation:

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

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 extremely 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 two things in the current context. First, it provides strong evidence that the moderately deep crawl we are proposing will capture the large majority of surfing behavior away from the seed sites. Consistent data from a wide variety of Web contexts provides a high degree of confidence that roughly 80% of searches will terminate before exceeding the depth of the crawl we perform. Even many searches that do exceed this depth will likely stay within the boundaries of our search set, given the small diameter of the Web. Second, the benefits of a deeper crawl appear to be modest. Huberman's work suggests that 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, downloading and analyzing 4.5 million Web sites for *each* of the 12 crawls would ultimately require more than 5 terabytes of disk storage.

4 Results

The six political topics that this paper examines are quite different from one another. Abortion, gun control, and capital punishment vary in the number and size of their advocacy groups, in their level of popular engagement, and in their relationship with formal governmental institutions. Web pages focusing on the President and on the U.S. Congress are presumably quite different both from each other and from pages which focus on a particular political issue. Then there is the general politics category—an area for which the Google seed set seems too broad, and the Yahoo seed set seems too narrow. And of course, the Google and Yahoo directory seed sets diverge significantly from each other.

Our research design introduces numerous sources of potential heterogeneity. The level

	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.

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 inlinks.

First, let us examine again 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, it bears repeating, is quite large—most weighed in at a little less than a quarter of a million pages. All told, we analyzed just shy of 3 million pages, not accounting for overlap. 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.

Still, out of the large number of pages crawled, only a small fraction were relevant to the given category. Again, previous research suggests that these crawls should have captured almost all of the content accessible from the Web's two most popular search tools (Huberman et al. (1998)). Abortion is, by many measures, the most divisive topic in domestic politics. It is the focus of much grass-roots political organizing, and millions of citizens have been mobilized on one side or another of the issue. Our research suggests that the universe of pages that are easily reachable using these methods is 10,000 to 12,000, smaller than some may have thought.

	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.

Table 1 suggests, too, that the SVM classifier is 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. Subjective coding of these sites suggests that most should be included in the positive set.

Two reasons suggest that these marginal sites pose little problem for our analysis. First, in general, these sites attract few inlinks, and as such they are unlikely to be a central part of the online community surrounding a given topic. Second, and most important, secondary analyses conducted with "unsure" sites included in the positive set found no substantive differences in the results detailed below. If anything, the reported results would be even stronger with their inclusion.

In several cases, the overlap between the Google and Yahoo seed sets was small. This was initially a source of some concern, even within our research group, that the communities crawled might not be directly comparable. Table 2, which shows substantial overlap between the Yahoo and Google positive sets, does much to alleviate those fears. It reinforces our conviction that the Yahoo and Google communities are closely linked, and provides a tangible demonstration of the small diameter of the Web.

Even the numbers above, however, don't tell the whole story. As we'll show in greater detail below, most of the pages in the positive set are relatively obscure, and contain only one or two inlinks. As one might expect, the least overlap occurs with pages which contain only one hyperlink path to them. Among the most heavily linked pages, the

	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.

overlap between the Yahoo and Google results is almost complete.

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 the whole. Of even greater interest than the size of these topical communities, however, is the way 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. Here 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. We have explained at length that the number of inlinks a site receives is a crucial measure of its visibility. Table 4 gets to the heart of the matter. The first column contains the

[7]It 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.

number of *sites* in each category which contain at least one positive *page*. For example, `abortionfacts.org` is an anti-abortion Web site maintained by the Heritage Foundation. `Abortionfacts.org` contains within it a number of 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 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 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. But in every case, these 50 sites account for the vast majority of inbound links.

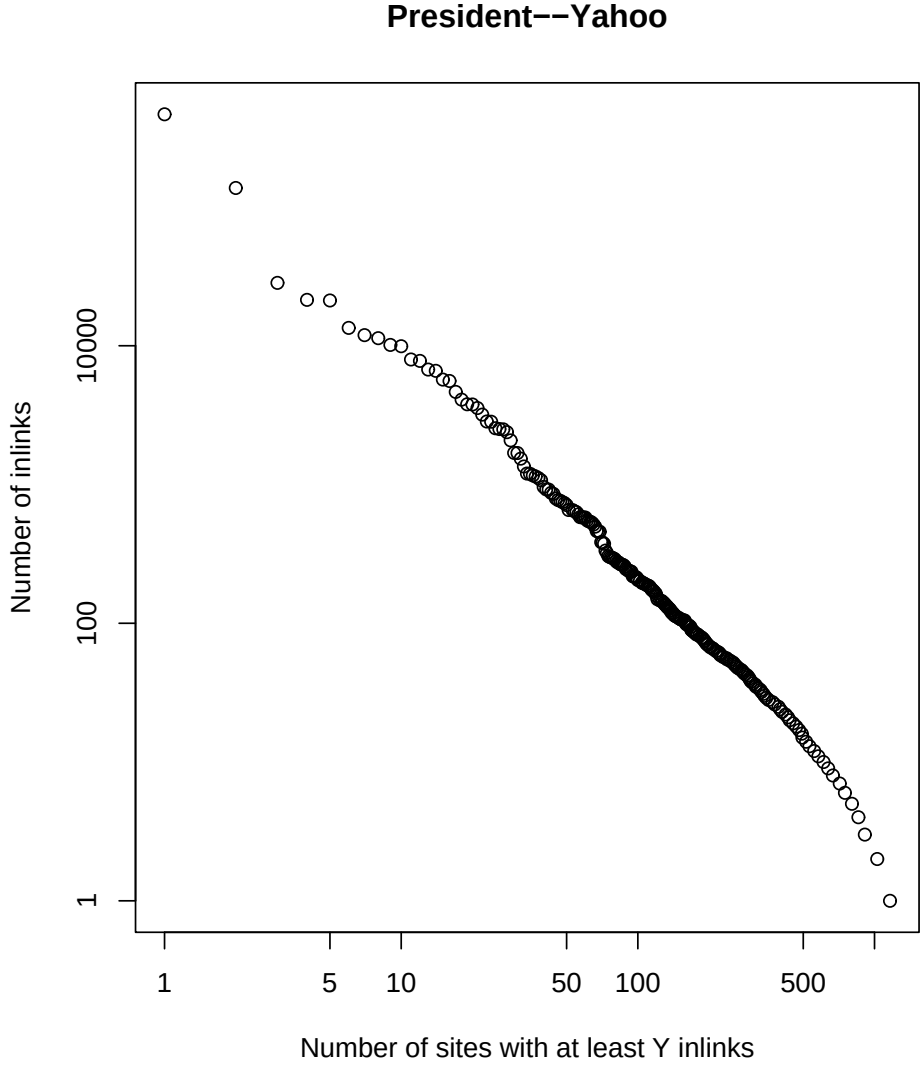

Figure 1: This chart shows the distribution of inbound hyperlinks for sites which focus on George W. Bush. Both axes are on a log scale. Note that the data form an almost perfect straight line—unmistakable evidence of a power law distribution.

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

To see exactly what the inter-community link structure looks like in practice, consider the figures below. Figures 1, 2, and 3 are designed to provide a representative cross-section of the dozen crawls. The first looks at sites which contain information on the President, the second looks at sites devoted to the death penalty, and the third examines sites dealing with general politics. Two of these examples are generated from Yahoo seed sets; the other is from Google. The Presidency community explored from the Yahoo seed set is the most concentrated community in the sample; Yahoo's general politics category is the least concentrated. The Google death penalty community is somewhere in the middle.

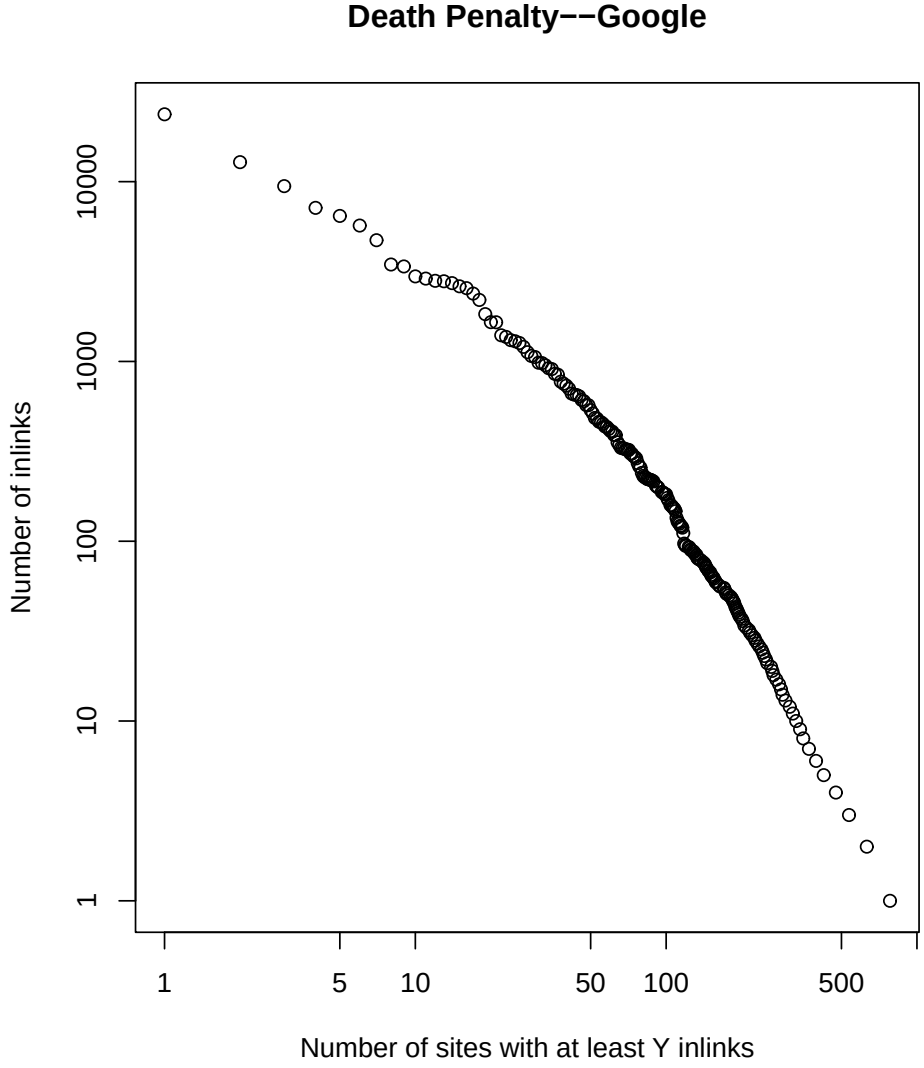

Figure 2: This figure illustrates the distribution of inlinks for sites focusing on the death penalty. Here again we see strong evidence of a power law distribution, although there is a slight upward bulge to the plotted data. Fitting a power law to to this data produces an R^2 of .9516—the second-lowest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

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 and Figure 3. 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. Both the death penalty community and the general politics community here deviate 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, however, it seems that power laws do an excellent job of characterizing link

[8]The 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).

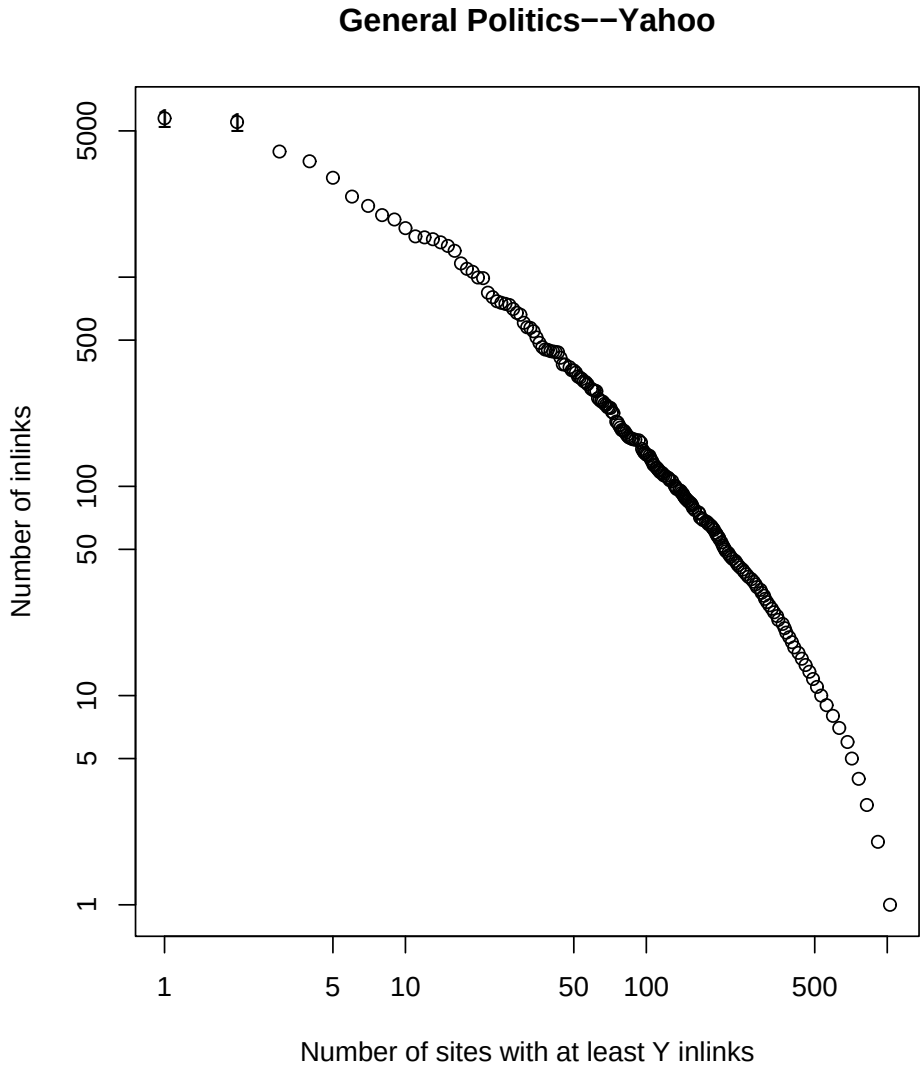

Figure 3: This chart shows the distribution of inlinks for general politics sites encountered crawling away from the Yahoo seed set. Note again both the general power law distribution and the slightly curvilinear shape.

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 of the number of links pointing to a given Web site (e.g. 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 distri-

	Coefficient ($-\alpha$)	Constant	R^2
Abortion (Yahoo)	-1.544	11.834	.9016
Abortion (Google)	-1.488	11.819	.9722
Death Penalty (Yahoo)	-1.684	12.007	.9766
Death Penalty (Google)	-1.958	13.960	.9516
Gun Control (Yahoo)	-1.458	11.650	.9612
Gun Control (Google)	-1.806	13.113	.9681
President (Yahoo)	-1.659	13.014	.9922
President (Google)	-1.705	13.285	.9746
U.S. Congress (Yahoo)	-1.909	13.239	.9706
U.S. Congress (Google)	-1.530	12.952	.9529
General Politics (Yahoo)	-1.252	10.583	.9562
General Politics (Google)	-1.454	13.536	.9770

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 coefficent 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 that the most popular site is predicted to receive.

bution 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. Also, as the figures above suggest, the power law model consistently predicts greater numbers of inlinks for the four or five most successful sites than we see in the data; to a much 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 serve as to dilute—at least slightly—the concentration of attention on the small number of successful sites.

On the whole, however, these qualifications do little to change the overall pattern of link distribution. 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.[9] The body of the data, in *every* community, adheres stubbornly to a power law, and omitting the 5 highest and lowest link values generally produces a near-perfect fit. Inlink distribution within political communities thus seems bound by powerful statistical regularities.

Whether online communities are better characterized by power laws or by some other

[9]By way of comparison, the relationship between the frequency and severity of wars, which is also power law distributed, produces an R^2 of .985 by the same measure (Cederman (2003)).

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 highly regular and extremely rapid 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. In examining the top 20 or even the top 50 sites across these dozen crawls, remarkably few sites had any hint of grass-roots flavor. There are, of course a few instances where a site run by a single individual or a (formerly) small group has become prominent. In the general politics category, several "blogs" have risen to prominence, such as Joshua Micah Marshall's `talkingpointsmemo.com`. A few individual gun-rights advocates are running prominent second amendment Web sites. And then there is `bushorchimp.com`, which compares at length the visage of the 43rd President of the United States with the faces of our closest simian relatives. Still, there is no doubt that 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 intriguing and important point that deserves emphasis: *the power law structure persists even if these collections of positive 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 of sites, 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-defined subgroups of sites. 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.

5 Googlearchy, and What It Means for Politics

The World Wide Web is the most diverse medium that has ever existed. The billions of pages that make up the Web cover an inconceivably vast array of topics. It is all the more surprising, then, that the motley group of political communities that we examine in this study all share a single organizing principle. In all of the communities we study, the number of inlinks that a site receives follows an extremely skewed distribution. The degree of concentration does vary. It may take 2 sites to account for 50% of the inbound links, or it may take 20; the difference between these two numbers is of course substantively significant. But in every case a minuscule fraction of sites account for most of the inbound links.

We term this organizational structure "googlearchy": the rule of the most heavily linked. Each category of political information we explore seems to be dominated by a handful of heavily-linked sites. All modern search engine algorithms—including those radically different from Google's PageRank—tend to return these most connected sites first (Ding et al. (2002)). Even users who rarely use search tools are affected by this pattern of linkage, as sites that receive many inbound links are (not surprisingly) far easier to find than sites that can be reached only by a few paths. Rather than radically decentralizing the dissemination of information, the prospect of googlearchy suggests that political information may remain highly concentrated even in the online world. To paraphrase Orwell, on the Web all sites are equal—but some sites are more equal than others.

Of course, the mere fact that inequality exists on the Web is not surprising. Few assumed, after all, that the Web would produce strict equality among its constituent sites, any more than they assumed that the *Walla Walla Union-Bulletin* would have the same number of readers as the *New York Times*. What deserves to be reemphasized, however, is the scale of the disparity. To continue the analogy, the community of online newspapers does have a few superstars, but globally it follows a log-normal distribution where most sites have a substantial number of links (Pennock et al. (2002)). In *all* of the political communities we study, the median site has only a single inbound link.

What does the structure of the Web mean for politics? Although it is clear that inlinks within political communities are highly skewed towards a few winners, discerning the true implications of this finding for politics is a more speculative—and thus more controversial—enterprise. Nonetheless, claims that the Web will make political information less concentrated remain common both inside and outside the academy, and are used to draw quite weighty conclusions. While many scholars have argued that public discourse is changing, we suggest that this change is taking place in different ways than many have assumed. It is worth exploring the degree to which which recent scholarship would need to be revised, or at least recast, if the structure of political information online were taken into account.

While our work permits many applications, it thus seems most immediately applicable to scholarship on the normative implications of public discourse in the digital age. We examine briefly three tightly intertwined sets of concerns: the potential for media balkanization, the Web's impact on democratic deliberation, and the ultimate consequences of the Web on the competence of ordinary citizens.

5.1 "Narrowcasting" and Balkanization

The first clear implication of our research is to revise our understanding of the Web's most basic characteristics. The Web, many have claimed, is not a mass medium. It supposedly represents a fundamental shift from broadcasting to "narrowcasting," where content is not produced for the general public but carefully tailored for a much smaller audience. Cass Sunnstein, for example, imagines 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" (Sunnstein (2001), p. 3). With citizens getting their political information from countless unreliable and polarizing sources, Sunnstein argues, the Web means that democratic discourse will become a balkanized mess. In a similar vein, Joesph Nye suggests that narrowcasting may erode the sense of community which undergirds the nation state (Karmark and Nye (2002), p.11).

Googlearchy suggests that the medium is not as narrow as much rhetoric suggests. Communities of political Web sites may be large and of general interest, or they may

appeal only to a few. In both cases, however, the number of focal sites is likely to be small. The number of prominent sites does not seem to scale up as the amount of interest in a topic increases. In many cases, these prominent sites serve as gatekeepers for entry into the rest of the community. Popular hubs often bring content on less-trafficked sites to wider notice. This phenomenon may help lessen inequalities of attention. However, it still grants a few successful sites a large degree of *de facto* editorial control.

Claims that the Web will lead to political fragmentation depend on a cliché: that the Web allows countless wackos with bizarre—and frightening—views to have political influence. Our research shows that this cliché misunderstands the real danger. Small groups on the political margins can and do use the Internet to organize themselves, sometimes more effectively than was possible a decade ago. Carol Swain's recent work on white supremacist groups documents one example of this phenomenon (Swain (2002)). As Swain's research shows, however, most of this online neo-nazi activity is channeled to a few prominent sites like `stormfront.org`.

Our findings, then, challenge us to rethink both the definition and the implications of "narrowcasting." The concern is not that the Internet allows countless fringe groups to spring up; rather, it is that the Internet funnels viewers with particular fringe concerns to *a few* Web sites. While commentators have spoken about the Web's decentralizing influence on political discourse, the examples they give often show *centralization* and pooling of resources among non-mainstream groups using the Web. To put it another way: it is not the countless wackos with Web sites that should give democratic theorists pause; it is the small number of heavily-linked wackos which are a potential problem. If our arguments about googlearchy are correct, the risks of political balkanization are different in both scale and character than has been previously understood.

5.2 Democratic Deliberation

The past two decades have seen a great deal of scholarly interest in "deliberative democracy"—in the claim that properly structured public debate can increase political legitimacy, build social capital, and transform the identity of citizens in positive ways (Gutmann and Thompson (1996); Habermas (1996); Mansbridge (1984)). As Cass Sunnstein makes clear, concerns about media balkanization come directly out of this literature. But delibera-

tive theorists' hopes and fears about the Web are broader than just media fragmentation. Changing our assumptions about the visibility of online content naturally changes our expectations about what online deliberation will produce.

One of the central goals of discourse theorists is to strengthen what Gutmann and Thompson term "middle democracy"—discussion of political issues by average citizens, not just academics and policy elites. Many observers have assumed that the Web would make it easier for citizens to participate in deliberative activities, and a number of academics have attempted to measure the impact of online deliberative activities (see, for example, Price and Cappella (Forthcoming)).

Googlearchy revises prognoses about the Web's impact on deliberation in two ways. Firs, it suggests that it is hard for all but a few "ordinary citizens" to post their views prominently—and, conversely, to read the views of other ordinary citizens, unless they are highlighted on a small number of prominent sites. Political speech posted online—particularly speech without the resources of a large organization behind it—is simply not easily accessible, because it is obscured by countless other Web pages. Googlearchy presents structural barriers to the expansion of middle democracy.

Second, the past decade has seen a number of specially designed, interactive deliberative forums created. Googlearchy suggests that a few sites—like the successful `ethepeople.org`—will dominate the market for this sort of discussion. Having the bulk of citizen attention concentrated on a small number of sites may create problems of scale, and difficulties in ensuring that voices are equitably represented. The costs of participating in online deliberation will likely demonstrate the same sort of stratification seen in the offline world.

5.3 Citizen Competence

Finally, assumptions about the visibility of online information are crucial for claims that the Web may make citizens more or less qualified to make collective decisions. The competence of ordinary citizens is a perennial question in politics, both within and outside the literature on deliberation, and there has been widespread speculation about how the Web might affect the answer. Of course, much hinges on how competent citizens *want* to be, and our research does not add to our understanding of citizen motivation. However, it

does tell us a lot about the information available to those citizens who do choose to seek out Web sites focused on politics. Here the news, on balance, is generally positive.

In the communities we studied, most of the sites at the top of the power law distribution are run by familiar names: established political interest groups, government entities, or (more rarely) private companies. Almost without exception, they are good sources of basic, reliable political information. These sites are unlikely to present political views outside the political mainstream. Partly as a consequence, this small set of sites represents a rich and easily accessible resource for citizens who want to make more informed political decisions.

The strong regularities in Web structure we describe may prove tyrannical at times. But one can also chose to see them as a collective social enterprise—indeed, as evidence of a meritocracy of the highest order. As scholars, we often informally assume that an article cited by many other often-cited articles is an important piece of research. Our research might suggest to democratic theorists that the Web is run by similar principles. The end result is that, in the communities we examine, at least *some* high-quality content is easy to find. We can imagine a Web structure organized along more egalitarian lines. We can also imagine that a less focused structure might be a mess to navigate.

Scholars have often made twin claims about the Web's effect on politics: that it would simultaneously lower the cost of political information and reduce inequality of attention. Seldom have they recognized that, for the user, these are competing goals. Googlearchy suggests that political communities have traded the latter for the former, lowering the cost of information by focusing only on a few worthwhile sources. Though this makes the medium less open than its technical standards allow, it may help improve the competence of ordinary citizens.

6 Conclusion

In the 1989 movie *Field of Dreams*, Kevin Costner plays an Iowa farmer whose life is changed when he begins to hear voices in his corn fields. The voices repeat a powerful, persistent message: "If you build it, they will come."

Early work on the political impact of information technology often seemed to be motivated by the same earnest faith as Kevin Costner's character. Technology, we were told,

would allow anyone—even ordinary citizens—to post their own views and political insights on a medium with world-wide reach. Thousands upon thousands of political Web sites would be created. Most importantly—and herein lay the leap of faith—the existence of this cornucopia of content would expand enormously the political information available to, *and used by*, citizens. If the Web was built, citizens would come.

In most respects, our expectations about the Net have changed much over the past decade. It is perhaps surprising, then, that much (though certainly not all) of both popular and scholarly discourse still repeats a *Field of Dreams* narrative. This sentiment crops up in scholars' contemporary dystopian visions of the Net, in critical public policy debates such as the recent FCC decision, occasionally even in the work of scholars engaged themselves in chronicling the social barriers to Web access and effective use. It is this persistent set of assumptions that this paper has set out to examine, challenge, and ultimately revise.

This is necessary in part, we have argued, because discussions about the diversity of information "available" online often end up confusing two very different things: retrievability and visibility. While the Web's technical architecture means in principle that any computer online can retrieve any Web page, not all pages are equally likely to be encountered. What is needed, then, is a reliable proxy for Web site visibility. Many lines of argument and a large body of computer science research suggest that the link structure surrounding political Web pages tells us much of what we want to know. We have presented user data that shows that the number of links to a site is highly correlated with the number of visits that site receives.

We have thus proposed and implemented a new research methodology that leverages new computer science techniques and powerful hardware to tackle the problem of Web site visibility more directly. The result is the first large-scale survey of the content and structure of online political information. In this study we downloaded and analyzed almost 3 million Web pages—a non-trivial fraction of the World Wide Web. In each of the topical areas studied—from abortion to the U.S. presidency, the U.S. Congress to gun control, general politics to the death penalty—inbound hyperlinks follow a highly-skewed, roughly power law distribution. In every case, the information environment is dominated by a few sites at the top. Moreover, subcommunities within these topics also seem to follow a power law distribution. The results thus seem robust not just across diverse communities,

but also within communities at different levels of analysis.

This paper, of course, is not a direct refutation of the claim that the Internet will increase the diversity of information sources used by citizens. Such claims are necessarily comparative. Evaluating them properly would require comprehensive data on concentration in offline media, which is beyond the scope of this paper. Moreover, though the power law structures we find online are hardly egalitarian, they may still prove to be more open than traditional media.

The point of this paper is therefore more modest. Instead of comparative claims, we have returned again and again to a single argument: that links between pages follow a consistent pattern, and that this pattern has systemic consequences for the visibility of individual Web sites. While we cannot evaluate whether the Web enjoys a comparative advantage in terms of content diversity, it is clear that profound inequalities among Web sites complicate both initial hopes for the medium and more than a few current characterizations of it. This research is a first step, both in the theoretical literature it brings to bear on these political science problems and in the methodology it implements. The data we present on Web site inequality, however, is remarkably strong.

There are insights in these results for both computer scientists and political scientists. First, this research contributes to an emerging computer science literature on the structure of the Web. Previous research has hinted that power law distributions online might be an artifact of aggregation, and that within many communities one should expect to find a skewed but nonetheless far more equal distribution of links. This study suggests that communities where most sites have substantial numbers of inlinks are the exception, not the rule. The communities that have previously been studied at length—public companies, universities, newspapers—are all unusual, in that they represent groups in which there is a high degree of mutual recognition among the actors (Pennock et al. (2002)). The online communities that we examine are more typical, in which even those who run these Web sites cannot have a complete tally of their colleagues and competitors.

While the structure of the Web may have seemed an abstract curiosity to some social scientists, the study suggests quite concrete impacts on American politics. The bottom line is that a small number of political Web sites receive more attention than all the rest combined. How we interpret this finding, of course, depends greatly on our expectations.

But whether this result is surprising or not, it is clear that in some ways the Web functions quite similarly to traditional media. Yes, almost anyone can put up a political Web site. But our research suggests that this is usually the online equivalent of hosting a talk show on public access television at 3:30 in the morning.

We argue that this research on the structure of the Web both enriches our understanding of the medium's impact on politics and sheds light on important public policy decisions. For example, googlearchy should allow those who worry that balkanization is the necessary consequence of narrowcasting to rest easier. On the contrary, our research suggests that political information online will remain concentrated; at the same time, however, highly focused link structures may also allow potentially worrisome centralization within fringe political communities.

In other areas, the impact of these links structures on mass politics seems similarly mixed. It has been the fervent hope of techno-optimists that the Web would expand citizen participation in civic life. In the context of the scholarly literature on democratic deliberation, googlearchy makes this task harder than many have realized; with so much of users' attention focused on a few sites, communities remain largely closed and online deliberation still faces difficulties in ensuring that all voices are heard. The good news, however, is that the structure of the Web may well be a boon for the competence of citizens who use it. Concentration of attention online makes at least some high-quality content easy to find. In the process, however, most political Web sites are doomed to obscurity.

Many scholars have talked about the Web's ability to lower the cost of information, drawing explicitly or implicitly on the work of Anthony Downs and Mancur Olson. But it is useful in this case to supplement Downs and Olson with the works of such thinkers as Herbert Simon or Ithiel de Sola Pool. Both noted rather urgently that it could be costly to have too much information as well as too little. Rephrasing concerns he originally voiced in the 1950's, Simon declared that "What information consumes is rather obvious: it consumes the attention of its recipients. Hence a wealth of information creates a poverty of attention, and a need to allocate that attention efficiently among the overabundance of information sources that might consume it" (Simon (1971)). Computers may offer us orders of magnitude more information than previous generations enjoyed; but human attention, it seems, is not a scalable resource. Though Simon's thoughts on this subject

predate the transistor, the Web demonstrates the consequences of a poverty of attention on a massive scale. "New media" may prove to be more open than the old, but it nonetheless still concentrates user attention on a few sources of political information.

Googlearchy raises many questions about the ultimate political implications of the Web. There seems little doubt, however, that power law structures are a central feature of online political information. Googlearchy is a phenomenon political scientists must examine more closely if they are to understand the political consequences of the information age.

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.

Barabasi, Albert-Lazlo. March 3, 2003. Personal communication.

Barber, Benjamin R. 1998. *A Passion for Democracy: American Essays.* Princeton, N.J.: Princeton University Press.

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.

Broder, Andrei, Ravi Kumar, Farzin Maghoul, Prabhakar Raghavan, Sridhar Rajagopalan, Raymie Stata, Andrew Tomkins and Janet Wiener. 2000. "Graph Structure in the Web." *Proceedings of The Ninth International World Wide Web Conference.*

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.

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.

Etzioni, Amitai. 1993. *The Spirit of Community*. New York: Crown Publishers.

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.

Gutmann, Amy and Dennis Thompson. 1996. *Democracy and Disagreement*. Cambridge, Mass.: Harvard University Press.

Habermas, Jurgen. 1996. *Between Facts and Norms: Contributions to a Discourse Theory of Law and Democracy*. Cambridge, Mass.: MIT Press.

Hargittai, Eszter. 2003. "How Wide A Web?: Inequalities in Accessing Information Online." Doctoral Dissertation. Department of Sociology: Princeton University, Princeton, NJ.

Hoffman, D. L. and T. P. Novak. 2000. "The Evolution of the Digital Divide: How Gaps in Internet Access May Affect Electronic Commerce." *Journal of Computer Mediated Communication* 5.

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. Number 1398 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.

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. Forthcoming. "Which Public Goods Are Endangered?: How Evolving Communications Technologies affect *The Logic of Collective Action.*" *Public Choice.*

Manjoo, Farhad. 2003. "Can the Web Beat Big Media?" *Salon.* May 21.

Mansbridge, Jane. 1984. *Beyond Adversary Democracy.* Chicago: University of Chicago Press.

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. 2003a. Nielsen NetRatings Search Engine Rankings. Technical Report. Search Engine Watch.
URL: *http://searchenginewatch.com/reports/netratings.html*

Nielsen-NetRatings. 2003b. United States: Top 25 Parent Companies. Technical Report. Search Engine Watch.

NTIA. 2000. Falling Through the Net: Toward Digital Inclusion. Technical Report. National Telecommunications and Information Administration.

NTIA. 2002. A Nation Online: How American's 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' Economique 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.

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*

Price, Vincent and Joseph N. Cappella. Forthcoming. *Online Deliberation and Democracy.* Chicago: University of Chicago Press.

Richardson, Lewis F. 1948. "Variation of the Frequency of Fatal Quarrels with Magnitude." *American Statistical Association* 43:523–546.

Simon, Herbert. 1971. "Designing Organizations for an Information Rich World." *Computers, Communications, and the Public Interest.*

Sunnstein, Cass. 2001. *Republic.com.* Princeton, N.J.: Princeton University Press.

Swain, Carol. 2002. *The New White Nationalism in America.* Cambridge: Cambridge University Press.

Vapnik, Vladimir. 1995. *The Nature of Statistical Learning Theory.* New York: Springer.

Zipf, George. 1949. *Human Behavior and the Principle of Least Effort.* Cambridge, MA: Addison-Wesley.

A Support Vector Machines

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,[10] 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. Many social scientists will be unfamiliar with their operation. However, 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 points of 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 which teach the SVM where to draw the appropriate decision boundary. The goal is to draw a boundary cleanly 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, the SVM is "tested," and newly encountered points can be classified by their position in this space. In our simple two-dimensional example, the SVM would assume that any new point above the line was a square, and any point

[10] For an accessible and widely-cited introduction to support vector machines, see Burges (1998).

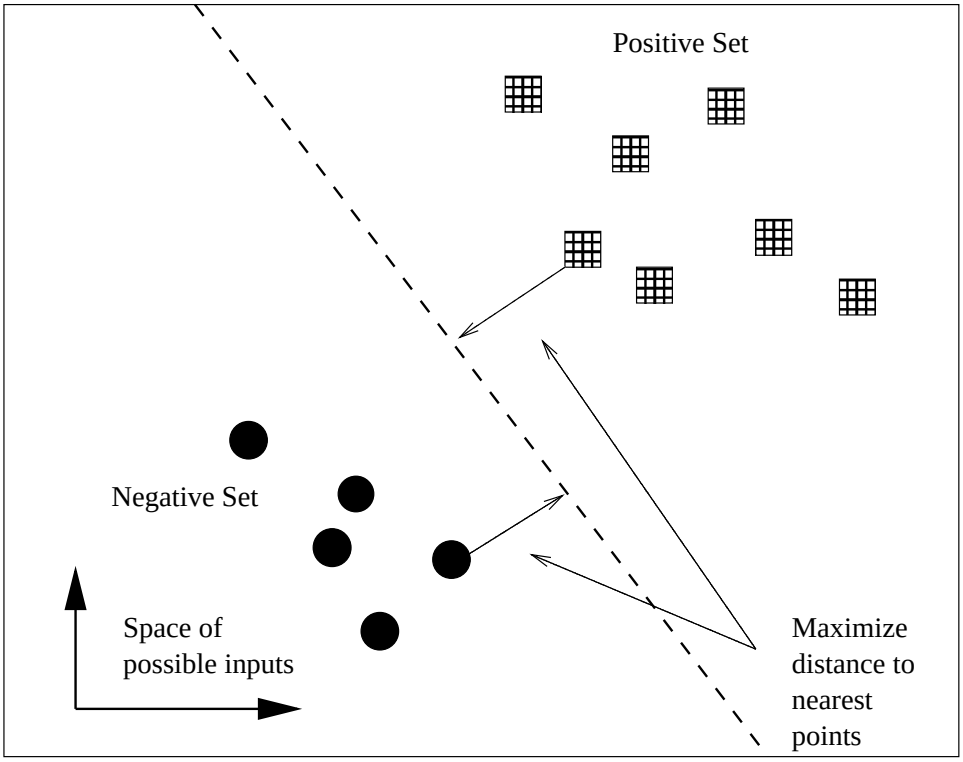

Figure 4: This figure shows a simple linear support vector machine. The boundary decision line is drawn to maximize the distance between itself and the *support vectors*, the points closest to the line. This example owes much to the explication of Platt (1998).

below the line was a circle.

For text classification, the text object is converted into a single point in a very high-dimensional space. In our analysis, the text object is the HTML document representing a particularly Web page. The HTML document is broken up into a series of *features*, which are either words or word pairs. Mathematically, 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 for them 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)). Traditionally, training of a support vector machine required solving a very large quadratic programming optimization problem. SMO greatly simplifies the computational demands by temporarily and sequentially fixing the values of the parameters to be estimated. Instead of a single very difficult problem, we solve a series of much smaller problems with analytic solutions.