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

Matthew Hindman† Kostas Tsioutsiouliklis‡ Judy A. Johnson§

March 31, 2003

Abstract

Many political scientists have assumed that the World Wide Web would lower the cost of political information and reduce inequalities of attention for those outside the political mainstream. However, computer scientists have consistently reported that the aggregate structure of the Web is anti-egalitarian; it seems to follow a "winners take all" power-law distribution, where a few successful sites receive the bulk of online traffic. In an attempt to reconcile these apparently disparate conclusions, this study undertakes a large-scale survey of the political content available online. The study involves iterative crawling away from political sites easily accessible through popular online search tools, and it uses sophisticated automated methods to categorize site content. We find that, in every category we examine, a tiny handful of Websites dominate. While this may lower the cost of finding at least some high-quality information on a given political topic, it greatly limits the impact of the vast majority of political Websites.

*Paper to be presented at the Annual Meeting of the Midwest Political Science Association, Chicago, IL.

†Ph.D. candidate, Department of Politics, Princeton University. (National Center for Digital Government, 104 Mt. Auburn St. Office 303, Cambridge MA 02138; `http://hindman.cc/`, `mhindman@princeton.edu`).

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

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

1 Introduction

Scholars disagree about what the Web means for politics. It has been aa decade since the release of Mosaic, the world's first graphical Web browser. Now more than 54% of the U.S. population uses the Web (NTIA (2002)). Online political activity seems to be common; according to the 2000 General Social Survey, 30% of Web users report seeking out political information online within the last 30 days. The Web has moved from a novelty to a mundane—but indispensable—part of the nation's economic and social life.

Still, as the impact of the Internet has become increasingly concrete, controversy has persists over its political implications. Some scholars believe that the Web will usher in a new era of participatory democracy (Castells (2000)). Others conclude that the Web will lead to increased isolation, balkanization, and the loss of a common political discourse(Sunnstein (2001); Barber (1998)). Still others offer a more nuanced analysis of how these new information technologies change the logic of collective action (Lupia and Sin (Forthcoming)).

This disagreement is curious—not because of its intensity, but because the scholars involved all share common assumptions that lead them to quite divergent conclusions. A. J. Liebling famously remarked that "Freedom of the press is guaranteed only to those that own one." Lupia and Sin are in good company when they argue, in essence, that the Web has made Liebling's aphorism obsolete:

> 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)).

Lupia and Sin's assessment of how the Web affects the information environment is widely shared. The remarkable openness of the Web's architecture has led many

scholars to make two assumptions. The first assumption is that the Web will generate a great deal of new, easily accessible content, thus lowering the cost of political information. The second assumption is that the Web will ameliorate inequalities of attention to views and information sources that are outside of the political mainstream. Though scholars have alternately celebrated the promise of the electronic town hall and worried that citizens will get their political news from unabomber.com, these two tenets have been viewed as self-evident features of the medium. While there is continuing concern about a potential "digital divide," the consensus seems to be that if everyone had equal access to the Web, and sufficient skills to use it, these twin promises of the Web would be fulfilled (DiMaggio et al. (2001)).

This paper challenges both of these assumptions. It argues, at length, that both the cost of political information online and the Web's ability to give voice to those on the margins is largely determined by the link structure of the Web. More specifically, we argue that the visibility of a site within an online sub-community is largely a function of the number of inbound hyperlinks it receives. If this is the case, however, it would seem to be bad news for the hopes of democratic theorists. Computer scientists have consistently shown that, for the Web as a whole, a small number of sites receive a hugely disproportionate share of inbound hyperlinks—and, partly as a consequence, a hugely disproportionate amount of traffic(Broder et al. (2000); Barabasi and Albert (1999); Kumar et al. (1999)). No one expected that every page on the Web would receive an equal share of attention. Still, the disparities that computer scientists have documented are staggering.

There remains a good deal of scholarly uncertainty about what this result means. The scholars who first described this result found it unexpected, and the best models to predict its formation leave important empirical features of the Web's link distribution unexplained. Moreover, there is debate about whether the global properties of the medium apply at the micro level, within clusters of topically related Web sites. Scholars who have examined a number of prominent subcommunities on the Web—

universities, newspapers, publicly listed companies—found that links within these groups followed very different, roughly log-normal or gaussian distributions(Pennock et al. (2002)). It remains an open question whether most online communities are governed by power laws, or whether the aggregate characteristics of the Web come from summing across a large number of subcommunities with skewed but unimodal distributions.

Still, the knowledge that the global structure of the Web is governed by a power law distribution suggests a very specific danger—a result we call "googlearchy." The Google search engine ranks results based on the number of inbound hyperlinks that a site receives. The suggestion is that this phenomenon may extend beyond just a single search tool, no matter how prominent—that a handful of heavily-linked sites may dominate political information on the Web, even for those who rarely use a search engine. Rather than "democratizing" the dissemination of information, the prospect of googlearchy suggests that citizens may continue to get their political information from only a few sources, even on the apparently limitless information vistas of cyberspace.

All of this raises crucial questions about the implications of the Web for politics. First of all, how much political information exists online for those with the most common political concerns? How is this information organized? Does it follow a power-law distribution? Or do political communities on the Web possess a different, and presumably more egalitarian, organizational scheme? How much of this information is truly easily accessible—truly low-cost—for the average citizen?

This paper attempts to answer these questions by applying recently developed techniques in computer science to networks of political Webpages. For the purposes of this paper, we focus on six different categories of political Websites. We use the most popular online search tools, Yahoo and Google, to create "seed sets" using their most highly-ranked results in these categories. We then crawl outward from each of these seed sites, using automated methods to classify newly encountered pages as

relevant to a given category of political content. Our techniques produce estimates of the total amount of accessible political content in each category. They also tell us a good deal about the structure of political information online—and about how egalitarian this new medium proves in practice.

2 Why Link Structure Matters for Democratic Theory

2.1 What *Is* Low-Cost Information Online?

The focus on the cost of political information exhibited by many scholars of the Internet ties them to a long tradition in political science—one that dates back to the work of Anthony Downs and Mancur Olson (Downs (1957); Olson (1965)). Both Olson and Downs present models of politics where gathering political information is a central cost—arguably *the* central cost—of political organization and participation. Theorists of the Internet have applied the logic of Downs and Olson in a quite straightforward fashion. The advent of the Internet, so the argument goes, has to lower the cost of gathering politically relevant information. And since the high price of gathering this information has been an important impediment to political engagement, cheaper information should lead to greater participation. End of story.

Though this argument is attractive, it does contain at least three problems. The first is that, for an argument that is based on economic assumptions, it tells us little about the elasticity of demand for political information. With nearly half a century of data suggesting that the level of interest in politics among the American electorate is low(Campbell et al. (1960); Miller and Shanks (1996)), it may be overly optimistic to assume that citizens *want* to be more engaged in politics, even if there was a "cheap" way to achieve this. A perhaps more plausible outcome is suggested by Markus Prior's investigations of the affect of cable news on overall news consumption.

Prior concludes that the amount of news most people wished to see was small; most viewers watched less news with the advent of CNN, though a small group of news junkies greatly increased their exposure (Prior (2002)).

Second, even if demand for cheap political information is high, we must consider how the Internet affects the *relative* attractiveness of group endeavors and various forms of political participation. Lupia and Sin explain that previously stable collective endeavors can be endangered by competition when information costs are dramatically lowered. Lupia and Sin argue persuasively that lower information costs can discourage, as well as enhance, collective action (Lupia and Sin (Forthcoming)).

Both of the above are important correctives to simplistic conclusions about the Web's impact on politics. But there is a third, and even more basic, problem with the narratives offered regarding politics and the Web. Recent computer science research on the structure of the Web challenges the two critical assumptions that were mentioned above: that the Web will decrease both the cost of political information and inequality of attention. In examining the ways that the hyperlinks that make up the Web are distributed, researchers have discovered that these 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 distribution over many orders of scale (Barabasi and Albert (1999)). To put it more exactly, the probability that a randomly selected Web page has K links is proportional to $K^{-\alpha}$ for large K. Empirical studies which 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)).[1]

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

[1]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 $\alpha = 2.4$.

pages) receive hardly any links at all.

It may not be immediately obvious that these empirical findings by computer scientists have anything at all to do with the concerns of those who are interested in what the Web means for democratic theory. But consider again what it means to have a piece of political information be "low cost." Low cost political information is, quite simply, information that is easily found. On the Web, content can be found in two ways. First, it can be discovered by surfing away from known sites; or second, it can be uncovered 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 the "cost" of political information on a site.

2.2 A Formal Model of Web Surfing

We know that the amount of Web traffic is highly correlated with the number of inbound hyperlinks connecting the site to the rest of the Web (Huberman et al. (1998)). For many, this result will be intuitive. The more links there are to a given site, the easier it is to find, and the more traffic it generates; it is much easier to navigate to `NYTimes.com`, for example, than it is to navigate to most personal Web pages. 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 dozens of less prominent sites.

Much of this understanding can be formalized and made more explicit. Consider the following simplified model of Web surfing behavior[2]. Let the "Web" be represented as a graph of N interconnected nodes $S_1 \ldots S_N$, each of which represent 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

[2]This model is derived from a model 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).

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.

Provided that every site in S is reachable from every other site in S, as $T \to \infty$ the proportion of traffic that accrues to a site S_j is directly proportional to the number of edges leading to the site from other sites in S. This is a somewhat strong assumption, however; 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 contingency, 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 which are heavily linked to, by other sites which are also heavily linked to, receive more than their share of Internet traffic.

To repeat: low cost 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 thought experiment above offers an insight into the former activity. It reinforces the conviction that the link structure of a site—and particularly the 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 "low cost" 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 most 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 the subterfuge? 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.* 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. One might think that a greater diversity of search engines would do much to ameliorate the problem. But in truth, the popularity contest dynamics associated with Google and PageRank are almost impossible 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 the relevancy

of results (Kleinberg (1999); Marendy (2001)). But Ding et al. show that despite the fact that the HITS approach is at the other end of the search engine spectrum from PageRank, it 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)). No matter what search engine is used, the (generally 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 cost 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. Social scientists would never assume that equality of opportunity in the economic sphere would result in an equal distribution of wealth. But many social scientists 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.

It is important to be clear: we are *not* suggesting that the number of inlinks to a site is the *sole* determinant of the cost of political information. There are, demonstrably, important exception to what we have been suggesting. For example, some citizens have extremely specific political concerns. Provided that search tools are adequate, and that these users are skilled enough to use search terms that reflect the specificity of their interests, the Web may indeed make political material for these small-scale topics available for a much lower cost in time and effort.

Other exceptions are important too. Websites may also offer different kinds of information or more easily navigable content than traditional media. The *New*

York Times site, like many, allows users to personalize content, and can even email subscribers when an item relating to a topic of interest appears —innovations that do lower the cost of political information even if the source of that information is unchanged. Or consider the example of an opponent of gun control who goes straight to `nra.org` without consulting a search tool. Offline information is routinely leveraged to find sites with in-depth information from sources already known and trusted.

Nonetheless, the potential limitations we've outlined above are profound. Sites that are found using offline information don't ameliorate inequalities of attention, they perpetuate it, introducing the biases of the old media into the new. Large numbers of the population either have quite general political interests, or cannot formulate their information requests with sufficient skill. For these users, an Internet where the distribution of links among political sites followed a power-law distribution would seem to offer only modest advantages over conventional media. A small number of sites would receive the lion's share of traffic, both that which originates from surfing away from known sites, and that which originates from online search tools. The bottom line is that content that is unpopular would remain more difficult—and thus higher-cost—to find.

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 a power-law distribution. Within specific categories of sites—for example, within all business homepages, or within all newspaper homepages—some researchers have found that the distribution of hyperlinks obeys a unimodal, roughly log-normal or gaussian distribution (Pennock et al. (2002)). However, these communities which have been found to deviate from the expected distribution have done so to differing degrees. It is unclear whether we should expect subcategories of political sites to be among them.

The ultimate conclusion is that the topology of the Internet may cause us to

question popular claims about the medium. Whether or not the Internet truly lowers the cost of retrieving political messages—and certainly the Web's potential to broaden the sources of political information used—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 Webpages. We solve this problem with the use of a support vector machine classifier. The SVM classifier offers two advantages. 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, and most important, the SVM classifier produces highly reliable categorizations.

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 computers scientists and learning theorists in recent years[3], 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

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

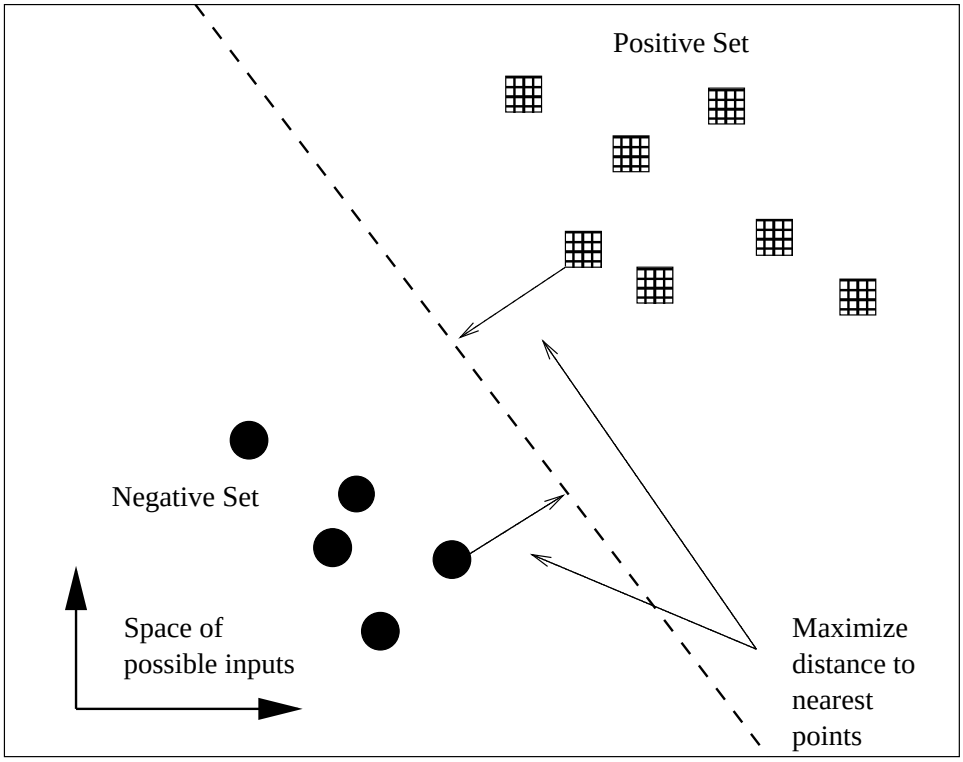

Figure 1: This figure shows a simple linear support vector machine. The boundary decision line is drawn to maximize the distance between itself and the *support vectors*, the points closest to the line. This example owes much to the explication of Platt (1998).

SVM's operate. In Figure 1, 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 circles 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 quite 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 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 features—for example, the phrase "United States"—occurs in the given page; otherwise the value is zero. One of the primary advantages of SVM's is that the difficulty of learning for them depends on the complexity of drawing the appropriate margin, and is largely independent of the dimensionality of the feature space.

For the purposes of this paper, we implement Platt's algorithm for sequential minimal optimization (SMO) in order to train our support vector machine (Platt (1998)). Traditionally, training of a support vector required solving a very large quadratic programming optimization problem. Platt 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.

The key point is that the support vector machine classifier proves reliable for our purposes. Most importantly, it produces very few false positives. Randomized human coding of sites shows that well over 99% 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 does 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 sets contain content about longstanding, controversial political issues: abortion, gun control, and capital punishment.

Seed set 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.[4]

[4]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

3.4 Surfer Behavior and Crawl Depth

For the purposes of this study we crawl each seed site to a depth of 4, three clicks away from our seed set. This depth was chosen for both theoretical and practical reasons. 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 Barabsi (1999)). 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 even with cutting-edge hardware. 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. have explained that page hits, like like inbound hyperlinks, follow a power-law distribution (Huberman et al. (1998)). But Huberman et al. also show that the number of links that a user will follow away from a starting Web site can be modeled extraordinarily well by an inverse Gaussian distribution. 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

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.

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 more than 75% of searches will terminate before exceeding the depth of the crawl we perform. Even many searches which 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 6%, 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 in a single would require more than 5 terrabytes 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. Webpages focusing on the President and on the U.S. Congress would seem quite different both from each other and from pages which focus on a particular political issue. And 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.

Our research design introduces numerous sources of heterogeneity. The level of consistency in our results, therefore, is all the more striking. All twelve of the crawls reveal communities of Web sites with similar organizing principles and similar

	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 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.

distributions of inlinks.

First, let us examine again the scope of the project. Table 1 lists the number of pages downloaded, as well of 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 millions pages. All told, we analyzed just shy of 3 million pages, not accounting for overlap. The size of the SVM positive set seems 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 of another of the issue. Still, our research suggest that the universe of easily reachable sites using

	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.

these methods is 10,000 to 12,000, smaller than some may have thought.

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 suggest that most should be included in the positive set.

Two reasons suggest that these marginal sites pose little problem for our analysis. First, most of 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 we may not have been crawling directly comparable communities. 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

	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.

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 sites, the 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 sites 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 to these positive sites 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

	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 is a given category.

that we have identified coherent communities of pages.[5]

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 accessibility. Table 4 gets to the heart of the matter. The first column contains the 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 is, not surprisingly, smaller than the total number of pages.

[5]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. However, eliminating cross-links between pages hosted on the same site has no substantive effect on the pattern of inlink distribution.

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.

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? Figures 2, 3, and 4 provide powerful evidence that the answer is yes. These three examples 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. Ultimately, however, the specific examples chosen make little difference. The inlink distribution was remarkably similar across all of 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 perfectly straight, downward sloping line. This is precisely what Figure 1 shows—a textbook power-law distribution. A very similar pattern is evident in Figure 2 and Figure 3. Here, though, the line formed by the data bulges outward slightly; the slope of the line gets steeper as the number of sites increases. In both cases, the disparity in links attracted is somewhat greater than fitting a power law distribution to the top half

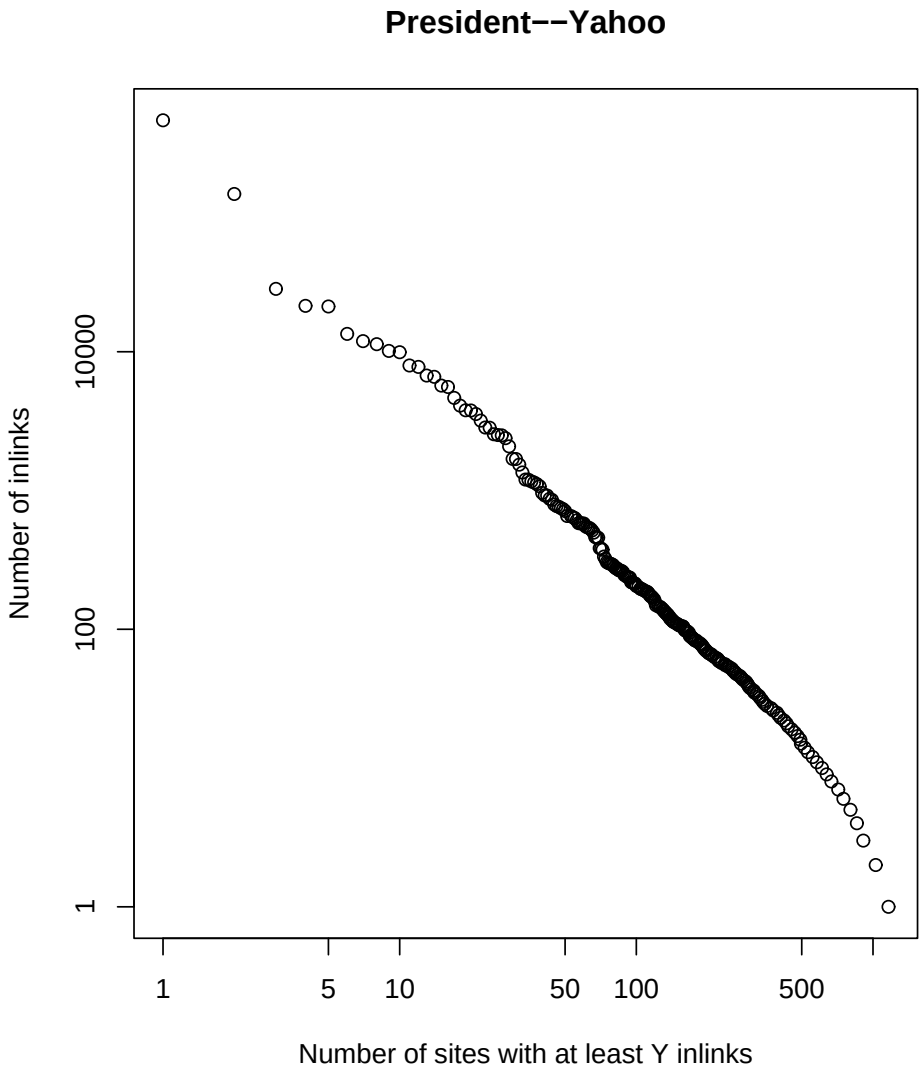

Figure 2: This chart shows the distribution of inbound hyperlinks for sites which focus on George W. Bush. Both axes are on a log scale. Note that the data form an almost perfect straight line—unmistakable evidence of a power-law distribution.

of the data would suggest.

Given the diversity both in seed sets and in the types of communities explored, these results are exceedingly strong—far stronger, in fact, than we had reason to expect. It seems a general property of political communities online that a small handful of sites at the top of the distribution receive more links than the rest of relevant sites put together.

There is one more related point to be made which does not show up in these charts and tables. 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

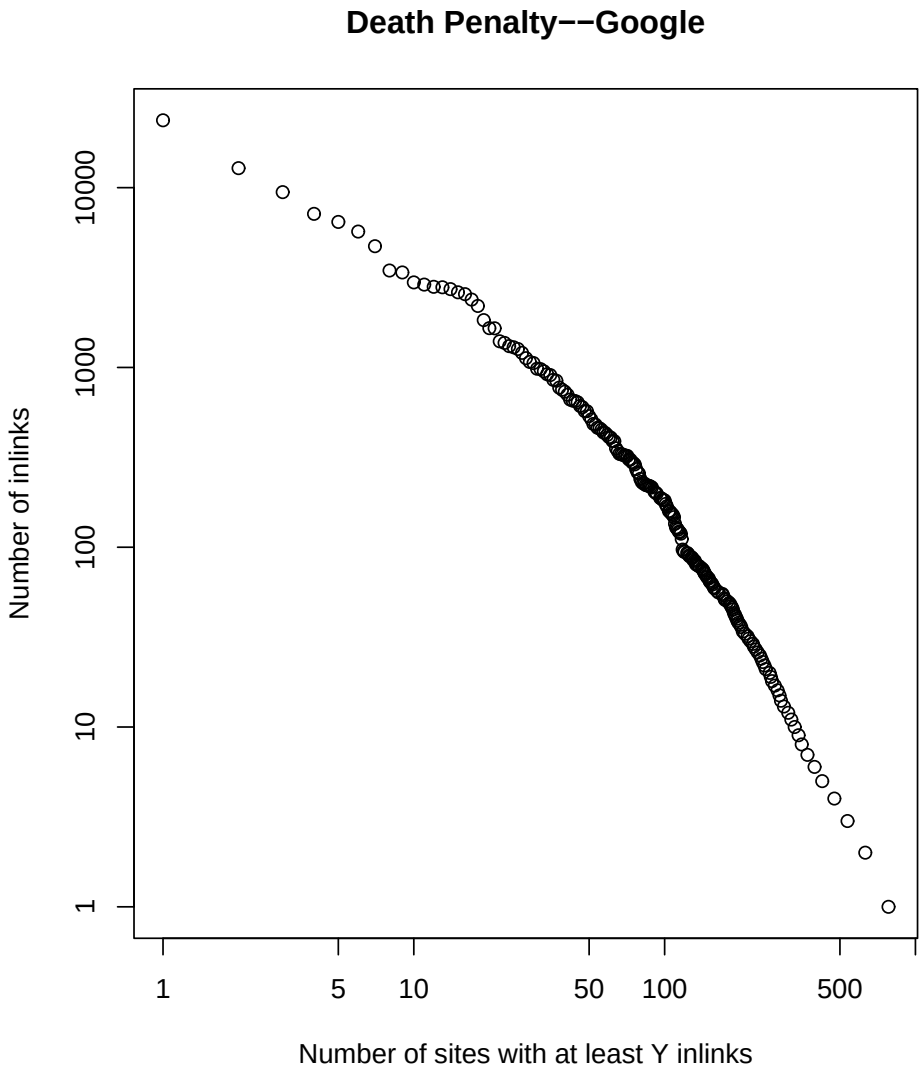

Figure 3: This figure illustrates the distribution of inlinks for sites focusing on the death penalty. Here again we see strong evidence a power-law distribution, although there is a slight upward bulge to the plotted data.

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 gun-rights advocates are running prominent second amendment Websites. 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.

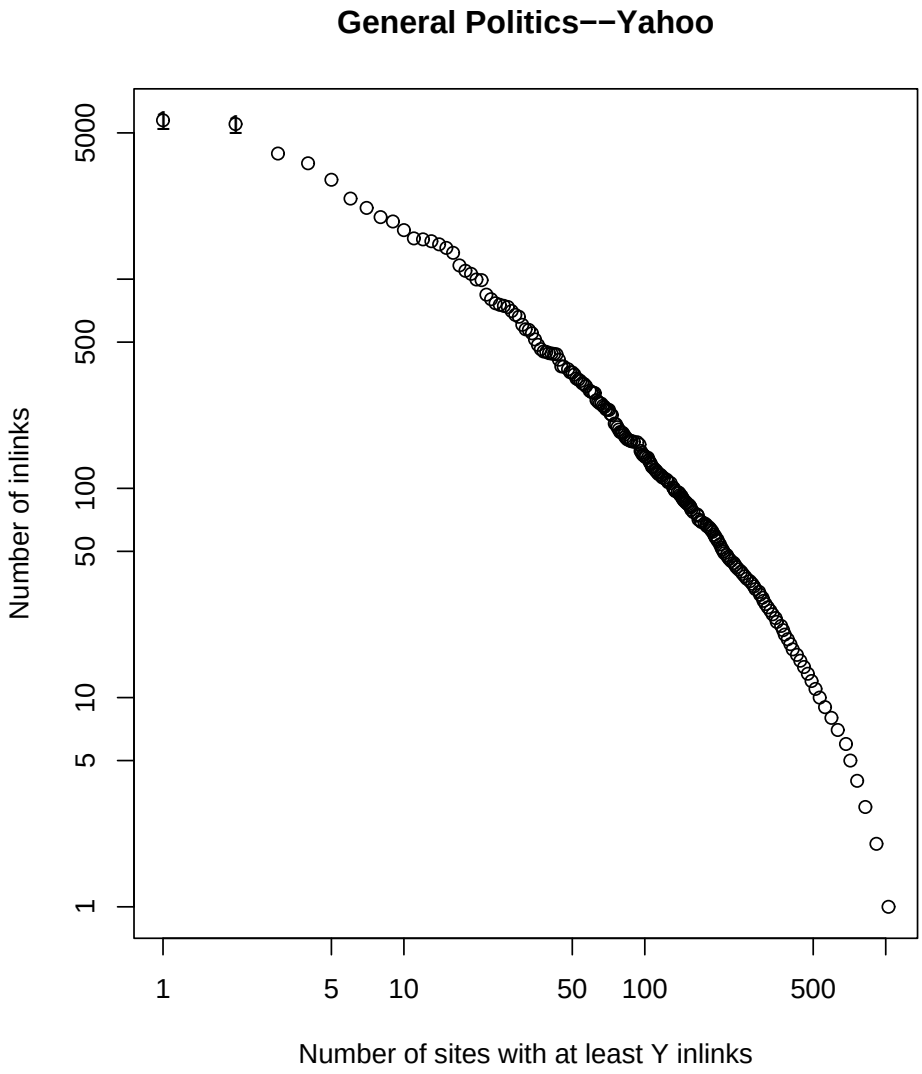

Figure 4: This chart shows the distribution of inlinks for general politics sites encountered crawling away from the Yahoo seed set. Note both the general power-law distribution and the slightly curvilinear shape.

5 Conclusion

The Web, it bears repeating, 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; but ironically, the very scale of this resource makes the Web 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 analysis, catalogued the ways that specific

groups have used the Internet to organize themselves. They have conducted surveys of Internet users, examining their demographic characteristics and broad patterns of usage. Still, the Web is so large that even surveys with hundreds of respondents tell us little about any single category of political information. The macro-level structure of political information has remained a mystery.

This paper takes a different approach, one that leverages new computer science techniques and powerful hardware to tackle the problem of scale 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 the process, 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 both for the visibility of individual Websites and for the flow of online information. 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—the distribution of inbound hyperlinks follows a power law distribution. In every case, the information environment is dominated by a few sites at the top.

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 unimodal distributions of inlinks are not common. The communities which have previously been studied at length—public companies, universities, newspapers—are all exceptional, 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 which we examine are unlikely to share that quality.

While the structure of the Web may be an abstract curiosity to social scientists,

the study suggests quite concrete impacts on American politics. Online search tools provide users with a point of entry into a community of relevant Websites. But given the small diameter of the Web, the shallow depth of most searches, and the fact that search engines such as Google are designed to mimic surfing behavior, this study should also provide us with a nearly complete survey of the sites that matter for a given political topic. The high degree of overlap between the results obtained with the Yahoo and Google seed sets—particularly among heavily-linked sites—underlines this point. Or to put it another way: any site which is more than three clicks away from *any* of the top 200 Google or Yahoo results on a given topic is definitely off the beaten track, and not likely to have any substantial impact on mass politics.

In cataloguing the sites that matter for politics, we have found that the number of important sites is small. Googlearchy—the rule of the most heavily linked—is the predominant feature of online political information. And at least in part, this surprising concentration of attention may be a good thing. Almost without exception, the most highly linked-to sites in a given category contain high quality information from credible (read: traditional) sources.

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.

Some of the results googlearchy suggests would seem to be good news for the democratic public. Scholars such as Cass Sunnstein and Benjamin Barber have

claimed that the Web will lead to a balkanization of political discourse, as individuals seek out countless Websites on the political margins. This research directly challenges that conclusion. At the same time, however, googlearchy provides substantial limits on the hopes of scholars who had invested the new medium with transformative power. The Web has proven to be more similar to traditional media than many originally thought. Yes, almost anyone can put up a political Website. For the vast majority of citizens, this activity is the online equivalent of having their own talk show on public access television at 3:30 in the morning.

In other areas of scholarship, googlearchy has similarly mixed consequences. If one wants to understand the role of the Web on a particular issue, for example, googlearchy may obviate the need to catalog thousands of sites, and allow scholars to credibly focus on a handful of the most popular and most heavily-linked sites. For the host of scholars who claim that the Web will lower the cost of political information, however, googlearchy urges greater caution. From the perspective of the user, the Web has indeed lowered the costs of many kinds of political information. Googlearchy suggests, however, that in the process most Websites have been doomed to obscurity.

Many scholars who have addressed this issue have approached it through the lens of Anthony Downs or Mancur Olson, and rightly so. 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.

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. The technical standards of the Internet still seem, to many, to hold enormous democratic potential. But open architecture is no guarantee of democratic outcomes. "New media," like the old, still displays extraordinary concentration of attention on a few sources of political information.

Whatever its consequences, googlearchy is a central feature of the online environment. It is a phenomenon political scientists must address if they are to understand the political consequences of the information age.

References

Albert, A., H. Jeong and A.-L. Barabsi. 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., R. Albert, H. Jeong and G. Bianconi. 2000. "Power-law distribution of the World Wide Web.".

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

Campbell, A., P. E. Converse, W. E. Miller and D. E. Stokes. 1960. *The American Voter.* New York: Wiley.

Castells, Manuel. 2000. *The Information Age: Economy, Society, Culture.* Oxford and Malden, Mass.: Blackwell.

Cortes, C. and V. Vapnik. 1995. "Support-Vector Networks." *Machine Learning* 20.

DiMaggio, Paul, Eszter Hargittai, W. Russell Neumann and John P. Robinson. 2001. "Social Implications of the Internet." *Annual Review of Sociology* 27:307–336.

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.

Downs, Anthony. 1957. *An Economic Theory of Democracy.* New York: Harper.

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.

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.".

Lupia, Arthur and Gisela Sin. Forthcoming. "Which Public Goods Are Endangered?: How Evolving Communications Technologies affect *The Logic of Collective Action.*" *Public Choice.*

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.".

Miller, Warren E. and J. Merrill Shanks. 1996. *The New American Voter.* Cambridge, Mass.: Harvard University Press.

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.
URL: *http://www.nielsennetratings.com*

NTIA. 2002. A Nation Online: How American's Are Expanding Their Use of the Internet. Technical Report. National Telecommunications and Information Administration.

Olson, Mancur. 1965. *The logic of collective action: public goods and the theory of groups.* 1 ed. Cambridge, Mass.: Harvard University Press.

Osuna, E., R. Freund and F. Girosi. 1997. "Improved training algorithm for support vector machines.".

Pandurangan, Gopal, Prabhakara Raghavan and Eli Upfal. 2002. Using PageRank to Characterize Web Structure. In *8th Annual International Computing and Combinatorics Conference (COCOON).*

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.".

Prior, Markus. 2002. "Efficient Choice, Inefficient Democracy?: The Implications of Cable and Internet Access for Political Knowledge and Voter Turnout." *Communications Policy and Information Technology: Promises, Problems, Prospects.*

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.

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