

IS 376: INFORMATION TECHNOLOGY & SOCIETY

Assigned Paper #3: Investigative Paper (100 points)

Due Date: Thursday, March 23, 2006

Your third paper assignment in IS 376 is an investigative paper concerning the alleged bias in search engines, using the attached article ("Googlearchy or Googlocracy?" by Menczer, Fortunato, Flammini, and Vespignani) as context. The focus in your paper will be to investigate **for yourself** whether search engines like Google and Yahoo exhibit a bias. Your investigation will include the following activities:

1. Submit the phrase *death penalty* to a search engine. Examine the top ten hits that the search engine produces and categorize each of those sites as either "favoring", "opposing", or "neutral". Then do the same thing with the top fifty hits produced by the search engine.
2. Submit the phrase *anti-abortion* to a search engine. Examine the top ten hits that the search engine produces and categorize each of those sites as either "positive", "negative", or "neutral". Then do the same thing with the top ten hits produced by the search engine when the phrase *pro-choice* is submitted.
3. Finally, submit the phrase *President Bush* to a search engine. Examine the top ten hits that the search engine produces and categorize each of those sites as either "positive", "negative", or "neutral". Then do the same thing with the top ten hits produced by the search engine when the phrase *George W. Bush* is submitted.

Once you have accumulated your statistics, report the results in an organized fashion in your paper and explain the results in the context of the attached article. Be sure to address the following issues:

- Were the *death penalty* results skewed towards "favorable" or "opposing"? If so, provide a feasible explanation for this fact.
- Was there a statistical difference between the *death penalty* results for ten hits and for fifty hits? If so, explain possible reasons for this difference.
- How did the *anti-abortion* and *pro-choice* statistics differ? Explain possible reasons for this difference.
- How did the *President Bush* and *George W. Bush* statistics differ? Explain possible reasons for this difference.
- Discuss whether the results of your investigation supported or contradicted the conclusions reached in the attached article.

Keep in mind that you are reporting on the results of your statistical analysis, **not** discussing your own political and social views concerning the issues about which you're making search engine submissions. Do **not** include remarks about your personal views on the death penalty, abortion, or the President. The emphasis in your paper should be on **search engine bias**.

This paper is required to have a minimum of 1000 words (approximately four double-spaced pages, with a 12-point font and one-inch margins). It must be word-processed, double-spaced, and legibly printed. The due date for this assignment is **Thursday, March 23, 2006, at 3:30 PM**. Late papers will **not** be accepted.

Googlearchy or Googlocracy?

By: Filippo Menczer, Santo Fortunato, Alessandro Flammini, and Alessandro Vespignani

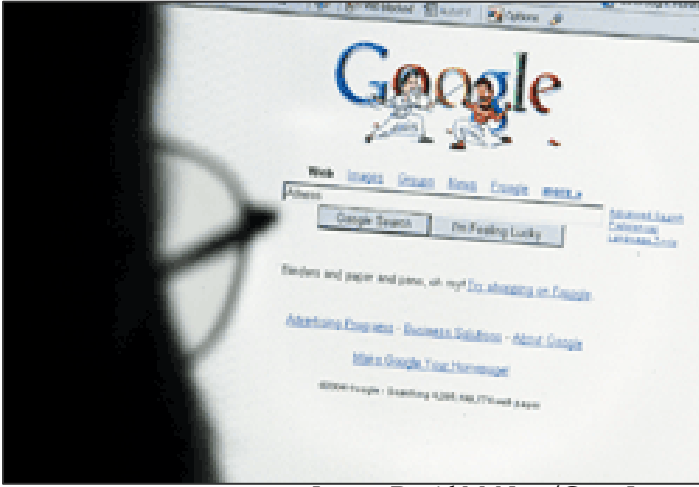


Image: David McNew/Getty Images

Search engines are our key to access information on the Web. Without search engines, we would easily become lost in cyberspace (as in the early days of the Web), so it is not surprising to see how heavily we rely on search engines as our information gateways. According to the Search Engine Round Table blog, Jay McCarthy, vice president of Web server log analysis company Websidestory, announced at the 2005 Search Engine Strategies Conference in Toronto that the number of page referrals from search engines has surpassed those from other pages. This means that people navigate the Web by searching more than by browsing.

The question of search engine bias then becomes a crucial one. What if search engines showed only certain types of information, or preferred certain sources? Imagine for example submitting the query *abortion* and finding only pro-life (or only pro-choice) sites in the first screen of hits. There are many kinds of potential bias—linguistic, political, cultural, commercial, and so on. The issue of bias resonates in the public debate on our growing dependence on search engines and on their social impact as gatekeepers of information. Is an information monopoly developing the same way as the software monopoly of the recent past? Is Google the next Microsoft? If search engines are the lens through which we see the world, transparency is a major concern, and any bias gets in the way. Our worries are heightened because search engines are secretive about their algorithms and, thus, their biases are subtle to detect.

In the midst of this debate, one kind of bias that has received much attention among technologists, as well as social and political scientists, is that in favor of "popular" sites. This stems from the PageRank algorithm, intro-

duced by the Google founders in 1998. All major search engines today use similar techniques to identify important or prestigious pages and bubble them to the top of the results. To a first approximation, PageRank attributes importance in proportion to the number of links that a page receives from other sites. The algorithm is a bit more sophisticated than that, but this approximation turns out to be pretty good on average (<http://arXiv.org/cs.IR/0511016>).

The notion of prestige based on link popularity is a proxy for other possible importance measures, such as traffic, expert judgment, and so on. Most people would agree that the use of prestige measures in ranking search results is a very good thing—indeed, it's the main reason why search engines work so well and have become so popular. Moreover, PageRank is designed to mimic the browsing behavior of Web users. In the absence of better assumptions, we imagine that people follow links at random. PageRank then estimates the traffic through each site. It seems, therefore, to be just the right criterion to rank sites. Why worry then?

To understand the potential danger of popularity bias, let us envision a scenario in which people search for information about the *minollo* (an imaginary animal). Imagine that there is an established site called *minollo-recipes.com* about the *minollo* and its culinary qualities. Further imagine a newly developed site called *save-the-minollo.org* that holds the view that the *minollo* is an endangered species and it should no longer be hunted. Now, suppose a student is assigned the homework of creating a Web page with a report on the *minollo*. The student will submit the query 'minollo' to a search engine and, for lack of time, browse only the top ten hits. Let's say that *minollo-recipes.com* is the fifth hit, while *save-the-minollo.org* is ranked 15th. The student will read the established site and write her report on *minollo recipes*. She will not read about the possible endangered status of the *minollo*. She will also diligently cite her source by adding a link from her new page to *minollo-recipes.com*.

As a result of this process, the more established site will have acquired a new link and increased its popularity (as measured by PageRank). The next time someone searches for information on the *minollo*, it will be more likely that the established site will be ranked even higher—fourth, say. The visibility of the less established site, on the other hand, will not increase. Now multiply this process by the billions of pages and links in the Web graph, and by the millions of queries handled by search engines every day. It is well known that the popularity of pages (measured

by their incoming links) has a distribution with a very long tail, in which a small fraction of very popular sites attract most of the links. This is the so-called rich-get-richer property of the Web. It seems reasonable to conclude that search engines amplify such inequality among Web sites by the vicious cycle described above: popular sites become even more popular, while new pages are less and less likely to ever be discovered.

While search engines do not make for a level playing field, their use partially mitigates the rich-get-richer nature of the Web, giving new sites an increased chance of being discovered.

People have called this presumed phenomenon by many names, including googlearchy (Hindman, M. et al., 2003. "Googlearchy: How a Few Heavily-Linked Sites Dominate Politics on the Web."). We have found many academic articles and blogs that discuss the phenomenon from technical, social, and political perspectives. Some scholars simply assume it, some present case studies that seem to support the idea, and a few technical papers attempt to prove and quantify the effect by indirect measures, as well as to develop remedies.

In a recent paper (<http://arxiv.org/cs.CY/0511005>), our group set out to quantify and model (predict) this effect by using empirical measurements that should allow us to directly gauge the effect. We followed the approach of Junghoo Cho at the University of California at Los Angeles and his collaborators and extended their analysis to connect the indegree (number of incoming links) of a site with the traffic to it. The two are linked by a chain of scaling relationships through PageRank, the rank of a search result page, and the probability that the user will click on a hit. There are three possible scenarios with clear scaling signatures. First, the googlearchy effect would generate a superlinear behavior of traffic as a function of indegree. Second, if search engines were popularity-neutral, i.e., if their ranking algorithms did not amplify page popularity beyond that determined by the Web graph structure, then the scaling would be linear, with traffic being simply proportional to popularity. Finally, a sublinear scaling would be the signature of a popularity mitigation effect by search engines.

To measure indegree we used the two major search engines, Google and Yahoo, and to measure traffic we used Alexa, the service that collects data from the users of its toolbar. These measures of indegree and traffic are the best available, but they are also notoriously noisy as some critics have observed ("Egalitarian Engines." *The Economist*, 19 Nov. 2005, p. 86). To deal with the noise, we sampled a very large number of random Web sites. We also employed logarithmic statistical analysis techniques, which are very robust to noise. This was possible because traffic and indegree span many orders of magnitude, and we only need to study their scaling relationship—not

precise individual values. We obtained remarkably consistent results using either Google or Yahoo data (which yield very different individual values) and by repeating our data collection over several months (during which the two search engines announced major upgrades of their collections).

Figure 1 shows the results of our analysis. The theoretical predictions are power laws, which appear as straight lines on the log-scale plot. The blue area represents the predictions corresponding to a googlearchy effect (super-linear scaling) while the line labeled "surfing model" represents the case in which search engines are neutral, as if users visited sites by surfing rather than searching. The empirical data do not fit a power law; but it is evident that traffic grows sublinearly with indegree. Contrary to our expectation, this result suggests that search engines actually have an egalitarian effect, directing more traffic than expected to less popular sites. Search engines thus appear to counteract the skewed distribution of links in the Web, directing some traffic toward sites that users would never visit if they were just surfing rather than searching. This egalitarian effect—which we could call googlocracy—is at odds with the arguments above. What gives?

To understand the observed googlocracy effect, we need to reconsider our model of how users visit sites as a result of their searches. The key factor that we have neglected in the original model is what kind of content users are interested in, and consequently what queries they submit. To develop a "semantically correct" model we must look at what people are searching for. We did this by analyzing a large query log from AltaVista containing almost 240 000 queries submitted by actual users. We then looked at how many results are returned by a search engine (we used Google) for these real queries. What we found is that like that of indegree, the statistical distribution of hit set size is also very skewed. Rarely are queries so general that the search engine returns a significant fraction of its collection; for example one in 1000 queries returns one tenth of the collection or more. The majority of queries return less than 30 000 hits (less than one page in a million from the collection), and for 4 percent of the queries there are only a screenful of results (10 or fewer hits).

When we take into account the semantic content of queries and how it affects searches, we get a semantically correct model that we can test by simulation. As shown in Figure 1, the prediction is remarkably accurate. The idea is that if a query returns few hits, then it is unlikely that globally popular pages will be included, notwithstanding their huge PageRank and indegree. On the other hand, new and less established sites have a higher chance to be relevant to specific queries, thus gaining visibility with respect to topics that have not (yet) been assimilated

by the major sources. Going back to our earlier example, if few people knew about the minollo, chances are that the student would have found fewer than ten hits, thus visiting both `minollo-recipes.com` and `save-the-minollo.org`. Since such specific queries are the majority, search engines have an egalitarian effect as they direct traffic to sites that would never be discovered by browsing alone.

While search engines do not make for a level playing field, their use partially mitigates the rich-get-richer nature of the Web, giving new sites an increased chance of being discovered, as long as they are about specific topics that match the interests of users. So it seems that cyberspace, for now, is more of a googlocracy than a googlearchy.

About the Authors

Filippo Menczer is an associate professor of informatics, computer science, and cognitive science at Indiana University, Bloomington. His research interests focus on intelligent systems for Web mining. Santo Fortunato is a postdoctoral research scholar at the Indiana University School of Informatics. His current research focuses on technological networks and the social dynamics of opinion formation. Alessandro Flammini is an assistant professor in the School of Informatics at Indiana University. His interests are mainly in the study of complex networks and in the physics of biopolymers. Alessandro Vespignani is a professor of informatics, cognitive science, and physics at Indiana University. He works on the theory of complex systems and networks.

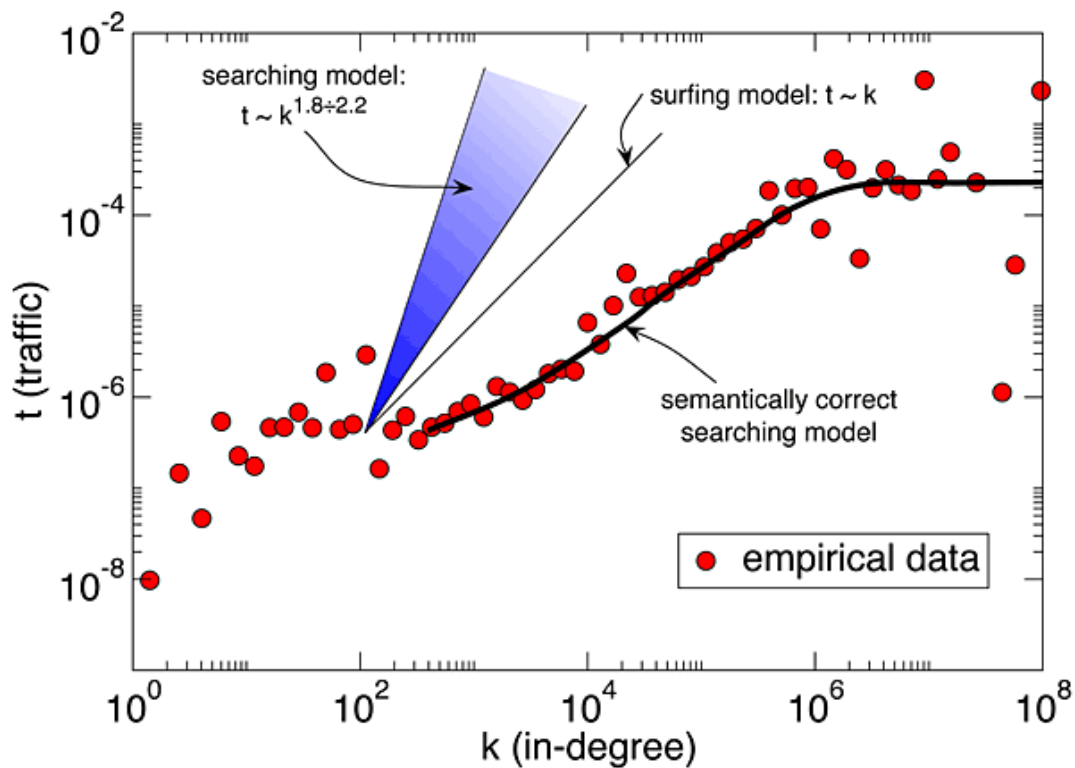


FIGURE 1: Scaling relationships between traffic and indegree of over 28 000 Web sites, about 2000 of which are the most popular sites and the rest are a random sample. Indegree is the number of incoming links to a particular site as reported by Yahoo, while traffic is the fraction of clicks by Alexa toolbar users that go to a particular site in a 3-month period. Data points are grouped into indegree bins of logarithmic size, and then the traffic is averaged among the sites in each bin. Note the increased noise for highest indegree values, due to the scarcity of data in that range—only a handful of sites like Google and Yahoo have that many incoming links.

FIGURE: MENCZER ET AL.