Web Spam, Propaganda and Trust

Panagiotis T. Metaxas
Wellesley College
Wellesley, MA 02481, USA
pmetaxas@wellesley.edu

Joseph DeStefano
College of the Holy Cross
Worcester, MA 01610, USA
joed@mathcs.holycross.edu

ABSTRACT

Web spamming, the practice of introducing artificial text and links into web pages to affect the results of searches, has been recognized as a major problem for search engines. It is also a serious problem for users because they are not aware of it and they tend to confuse trusting the search engine with trusting the results of a search [16]. The parallels between web spamming on the internet and propaganda in the real world suggest that we can use anti-propaganda techniques to educate users and develop tools to help them evaluate the reliability of the information they find online.

In this paper, we first analyze the effects that web spam has on the evolution of the search engines and their relationship to propagandistic techniques in society. Then, we examine the neighborhoods of untrustworthy sites, finding that a dense biconnected component (BCCs) containing the site provide a reasonable *trust neighborhood* that has parallels in social network theory. The fact that spammers employ propagandistic techniques enables us to design a heuristic that follows anti-propagandistic practices in order to recognize a spamming network. In society, recognition of an untrustworthy message (in the opinion of a particular person or other social entity) is a reason for questioning the entities that recommend the message. Entities that are found to strongly support more untrustworthy messages become untrustworthy themselves. So, social distrust is propagated backwards for a number of steps. Our heuristic simulates this behavior on the trust neighborhood of a spammer.

In our experiments, we examined trust neighborhoods of web sites, both trustworthy and not. Our findings suggest that spamming networks can be reliably recognized from their relationship to a single untrustworthy starting point by backward propagation of distrust. Further, nodes involved in a spamming network can be divided into two groups: those that have content similar to the starting site (aka "link farms"), and those that have dissimilar content (aka "mutual admiration societies"). Our tool explores thousands of nodes within minutes and could be deployed at the browser-level, making it possible to resolve the moral question of who should be making the decision of weeding out spammers in favor of the end user.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval; H.3.m [**Information Storage and Retrieval**]: Miscellaneous

General Terms

Algorithms, Experimentation, Social Networks, Propaganda, Trust

Keywords

search, Web graph, link structure, PageRank, HITS, Web spam

WWW2005, May 10–14, 2005, Chiba, Japan.

1. INTRODUCTION

The web has changed the way we inform and get informed. Every organization has a web site and people are increasingly comfortable accessing it for information for any question they may have. The exploding size of the web necessitated the development of search engines and web directories. Most people with online access use a search engine to get informed and make decisions that may have medical, financial, cultural, political, security or other important implications [10, 37, 23, 29]. Moreover, 85% of the time, people do not look past the first ten results returned by the search engine [35]. Given this, it is not surprising that anyone with a web presence struggles for a place in the top ten positions of relevant web search results. The importance of the top-10 placement has given birth to a new industry, which claims to sell know-how for prominent placement in search results and includes companies, publications, and even conferences. Some of them are willing to bend the truth in order to fool the search engines and their customers, by creating web pages containing web spam.

Web spamming is the practice of manipulating web pages in order to cause search engines to rank some web pages higher than they would without any manipulation.[1] The motive is usually commercial, but can also be political, or religious.

The creators of web spam are often specialized companies selling their expertise as a service, but can also be the web masters of the companies and organizations that would be their customers. Spammers attack search engines through text and link manipulations [22, 18]:

[1]We should mention here that there is not a complete agreement on the definition of web spam among authors, which leads to some confusion. Moreover, to people unfamiliar with web spam, the term is mistaken for email spam. A more descriptive name for it would be "search engine ranking manipulation" or "adversarial information retrieval".

- **Text spam**: This includes excessively repeating text and/or adding irrelevant text on the page that will cause incorrect calculation of page relevance; adding misleading meta-keywords or irrelevant "anchor text" that will cause incorrect application of rank heuristics.

- **Link spam**: This technique aims to change the perceived structure of the webgraph in order to cause incorrect calculation of page reputation. Such examples are the so-called "link-farms", "mutual admiration societies", page "awards", domain flooding (plethora of domains that re-direct to a target site), etc.

Both kinds of spam aim to boost the ranking of spammed web pages. Sometimes **cloaking** is included as a third spamming technique [22, 19]. Cloaking aims to serve different pages to search engine robots and to web browsers (users). These pages could be created statically or dynamically. Static pages, for example, may employ hidden links and/or hidden text with colors or small font sizes noticeable by a crawler but not by a human. Dynamic pages might change content on the fly depending on the visitor, fake the clickstream or query stream, submit millions of pages to "add-URL" forms of search engines, etc. We consider the false links and text themselves to be the spam, while, strictly speaking, cloaking is a tool that helps spammers hide their attacks.

Since anyone can be an author on the web, these practices have naturally created a question of *information reliability*. An audience used to trusting the written word of newspapers and books is unable, unprepared or unwilling to think critically about the information obtained from the web. In a recent study [16] we found that while college students regard the web as a primary source of information, many do not check more than a single source, and have trouble recognizing trustworthy sources online. In particular, two out of three students are consistently unable to differentiate between facts and advertising claims, even "infomercials." At the same time, they have considerable confidence in their abilities to distinguish trustworthy sites from non-trustworthy ones, especially when they feel technically competent. We have no reason to believe that the general public will perform any better than well-educated students. In fact, a recent analysis of internet related fraud by a major Wall Street law firm [10] put the blame squarely on the investors for the success of stock fraud cases.

One of the reasons behind the users' difficulty to distinguish trustworthy from untrustworthy information comes from the success that both search engines and spammers have enjoyed in the last decade. Users have come to trust search engines as a means of finding information, and spammers have successfully managed to get them to transfer that trust to the results of the search. There is clearly a need for education of the users, so that people develop a healthy suspicion of unverified search results. Beyond that, though, there is a need for browser-level tools that will help the user move from suspicion to decision in determining which sites to trust.

From their side, the search engines have struggled to deliver spam-free results and have developed sophisticated search result ranking strategies. Two such ranking strategies that have received major attention are the PageRank [6, 2] and HITS [26] algorithms. Achieving high PageRank has become a sort of obsession for many companies' IT departments, and the *raison d'être* of spamming companies. Some estimates indicate that at least 8% of all pages indexed is spam [12] while experts consider web spamming the single most difficult challenge web searching is facing today [22].

In this paper we first examine the reasons web spamming has been so successful and its relationship to social propaganda. Then, we develop heuristics that are able to recognize web neighborhoods, especially untrustworthy ones. We present experimental results that show considerable success in recognizing spamming neighborhoods. Finally, we discuss what we believe should be a frame for the long-term approach to web spam.

2. THE WEBGRAPH AS A SOCIAL NETWORK

The web is typically represented by a directed graph [8]. The nodes in the webgraph are the pages (or sites) that reside on servers on the internet. Arcs correspond to hyperlinks that appear on web pages (or sites). The theory of social networks [38] also uses directed graphs to represent relationships between social entities. The nodes (called "actors") correspond to social entities (e.g., people, institutions, ideas). Arcs (called "ties") correspond to social relations between the entities they connect (e.g., has influence on, knows, trusts).

This connection is more than a similarity in descriptions. The web itself is a social creation, and both PageRank and HITS are socially inspired algorithms [6, 26]. Socially inspired systems are subject to socially inspired attacks, however. Not surprisingly then, the theory of propaganda [28] can provide intuition into the dynamics of the web.

For example, PageRank is based on the assumption that the reputation of an entity (a web site in this case) can be measured as a function of both the number and reputation of other entities recommending it. A link to a web page is counted as a "vote of confidence" to this web site, and in turn, the reputation of a page is divided among those it is recommending [6]. Since HTML does not provide for "positive" and "negative" links, all links are taken as positive. This is not always true, but is considered a reasonable assumption.

More importantly, there is also the implicit assumption is that hyperlink "voting" is taking place independently, without prior agreement or central control. Spammers, like social propagandists, are groups of sites that are able to gather a large number of such "votes of confidence" by design, thus breaking the assumption of independence in a hyperlink. Search engines consider such moves spam, and would like to restrict it, but there can be no algorithm that can recognize spamming sites automatically based on graph isomorphism [5].

3. EVOLUTION OF THE SEARCH ENGINES

In the early 90's, when the web numbered just a few million servers, the **first generation** search engines were ranking search results using classic information retrieval techniques: the more rare words two documents share, the more similar they are considered to be [34, 21]. A search query Q is simply a short document and the results of a search for Q are ranked according to their (normalized) similarity to the query which was treated as the value of the page.

The first attack to this "$tf.idf$ ranking," as it is known, came from within the search engines. Around 1995, search

engines started selling search keywords to advertisers as a way of generating revenue: If a search query contained a "sold" keyword, the results would include targeted advertisement and a higher ranking for the link to the sponsor's web site. This is the first time we have a socially inspired ranking, which follows marketing practices of the real world.

Mixing search results with paid advertisement raised serious ethical questions, but also showed the way to financial profits to spammers who started their own attacks by creating pages containing many rare keywords to obtain a higher ranking score. In terms of propaganda theory, the spammers employed a variation of the technique of *glittering generalities* to confuse the first generation search engines [28, 47]. The propagandist associates one or more suggestive words without evidence to alter the conceived value of a person or idea. To avoid spammers (and public embarrassment from the keyword selling practice), search engines would keep secret their exact ranking algorithm. Secrecy is no defense, however, since secret rules can be figured out by experimentation and reverse engineering (e.g., [33, 30]).

Second generation search engines started employing more sophisticated ranking techniques in an effort to nullify the effects of glittering generalities. One of the more successful ones was based on the "link voting principle": Each web site s has value equal to its "popularity", which is influenced by the set B_s of sites pointing to site s. Lycos became the champion of this ranking technique and had its own popularity skyrocket around 1996 [31]. Doing so, it was also distancing itself from the ethical questions introduced by combining advertising with ranking.

Unfortunately, this ranking method did not succeed in stopping spammers either. Spammers started creating clusters of interconnected web sites that had identical or similar contents with the site they were promoting, which subsequently became known as "link farms" (LF). The link voting principle was socially inspired, so spammers used the well known propagandistic method of *bandwagon* to circumvent it [28, 105]. Using this technique, the propagandist is promoting the impression of a high degree of recommendation by inter-linking many internally controlled sites that will eventually all share high ranking.

The introduction of PageRank in 1998 was a major development for search engines, because it seemed to provide a more sophisticated anti-spamming solution to the bandwagon technique. Under PageRank, not every link contributes equally to the reputation of a page. Instead, links from highly reputable pages contribute much higher than links from other sites. That way, the site networks developed by spammers would not influence much their PageRank, and Google became the search engine of choice. A page p has value equal to its reputation $R(p)$ which is calculated as the sum of fractions of the reputations of the set B_p of pages pointing to p. HITS is another socially-inspired ranking which has also received a lot of attention [26]. The HITS algorithm divides the sites related to a query between "hubs" and "authorities". Hubs are sites that contain many links to authorities, while authorities are sites pointed to by the hubs. (This circular definition can be resolved.)

PageRank and HITS marked the development of the **third generation**.[2] Unfortunately, spammers have again found ways of circumventing PageRank. In PageRank, a page enjoys some "absolute reputation", that is, its reputation is not restricted on some particular issue. So, spammers develop sites with expertise on irrelevant subjects, and they justifiably acquire high ranking on their expert sites. Then they interlink their networked sites with the expert sites, creating what is called a "mutual admiration society" (MAS), causing all sites to share a higher PageRank and the search engine is fooled. This is the well-known propagandistic technique known as *testimonials*, where well known people (entertainers, public figures, etc.) offer their opinion on issues about which they are not experts [28, 74]. HITS has also shown to be highly spammable by this technique [25] due to the fact that its effectiveness depends on the accuracy of the initial neighborhood calculation.

The table below summarizes our findings for the first three generation of search engines and the correspondence between web spam and social propaganda.

SE	Ranking	Spamming	Propaganda
1st Gen	Doc Similarity	keyword stuffing	glittering generalities
2nd Gen	+ Site popularity	+ link farms	+ bandwagon
3rd Gen	+ Page reputation	+ mutual admiration soc.	+ testimonials

Web search corporations are reportedly busy developing the engines of the next generation [7]. The new search engines hope to be able to recognize the need behind the query of the user. Given the success the spammers have enjoyed so far, one wonders how will they spam the fourth generation engines. Is it possible to create a ranking that is not spammable? Put another way, can the web as a social space be free of propaganda? Seen in this light, it appears that we are trying to create in cyberspace what human societies have not succeeded in creating in their social space. However, as in society, we can learn to live successfully with propaganda, given appropriate education and technology.

4. EXPLORING WEB NEIGHBORHOODS

Since spammers employ propagandistic techniques, as we have argued above, it makes sense to design anti-propagandistic methods for defending against them. These methods need to be user-guided. Propaganda, after all, [11], does not always have a negative connotation. Advertisement is a form of propaganda that we have all learned to live with. The "art of persuasion" is objectionable mainly when it is used to promote an untrustworthy message according to the receiver's opinion. When such an untrustworthy message is detected, it becomes a reason for us to reconsider the messenger. Messengers who strongly support an untrustworthy message become untrustworthy themselves. This process is selectively repeated for a few steps, propagating the distrust of the original back to those who show support for it. The results of this process become part of the user's belief system and are used to filter future information.

Propagation of distrust contrasts with the propagation of trust in that it progresses *backwards* through the graph, i.e. a page's reliability *decreases* if it links to untrustworthy sites. Current algorithms for ranking pages propagate trust

[2][7] considers the search engines in our 2nd and 3rd generation to be in the same group. We believe that both the ranking and attack methods puts them in different categories.

forwards, i.e., a page's rank is increased if a trustworthy site links to it. The heuristic in this paper focuses solely on distrust, but in future work we plan to investigate the combination of the two.

Following the social process above, we design an algorithm that follows anti-propagandistic practices in order to recognize a spamming network. Our algorithm takes as input a page that the user determined to be untrustworthy. This page could have come to the user through web search results or via the recommendation of some trusted associate (e.g., a society that the user belongs to).

- In the first stage, our algorithm follows for a few steps the back links of the authority site that contains this page. Thus, we create a portion of the web graph that supports the starting site (its "trust graph"). In the process we also sample the contents of the sites in this subgraph to determine their similarity to the starting site's contents.

- In the second stage the trust graph's structure and contents are analyzed. The subgraph that heavily supports the starting site becomes suspicious for spamming. In our experiments we found that the biconnected component (BCC) of the graph that includes s is such an appropriate subgraph.

  We divide the sites in the BCC into two groups: Those that have contents similar to s, and those that do not. The former are considered members of a link farm, while the latter are considered members of a mutual admiration society. The members of the link farm are discredited, and the members of the MAS are downgraded (although this result could perhaps be user-selectable, for a less conservative filtering).

5. EXPERIMENTAL RESULTS

We define the *h-trust neighborhood* of some site S as the largest biconnected component containing S of the graph composed of the sites that are no more than h links away from S. In our experiments, we examined 3-trust neighborhoods of web sites, both trustworthy and not. We already know that the neighborhoods of spammers and non-spammers cannot be distinguished simply on graph-theoretic terms [14, 5]. They can, however, be recognized from an untrustworthy starting point by backward propagation of distrust.

To evaluate the trustworthiness of each site we had an evaluator look at the sites of the BCC. A site was then determined to be either Trustworthy, Untrustworthy, or Non-determined. The last category includes a variety of sites for which the evaluator could not make a clear determination due to the language used in the site, the subject matter, or the fact that a Blog or Directory can not fall simply into one of the U/T categories.

It is perhaps valuable to reiterate here the fact that we are considering trustworthiness to be a personal decision, not an absolute quality of a site. One person's gospel is another's political propaganda, and our goal is to find tools that help individuals make more informed decisions about the quality of the information they find on the web.

In selecting the theme to evaluate, we considered a number of commercial, political, medical and financial issues. For the purpose of this paper we focused on a subject that, according to the evaluator's opinion, was clearly untrustworthy. In particular, the evaluator decided that he does not believe that a product for $39.95 would significantly increase muscle mass without increased exercise, decrease fat without change in diet or habits, enhance sexual performance, increase the good cholesterol while decreasing the bad, regrow hair, decrease blood pressure, remove wrinkles, and increase memory retention. In our experiments, we examined the neighborhoods six such sites, as well as two sites judged to be trustworthy, labeled below as U-1 to U-6 and T-1 to T-2, respectively.

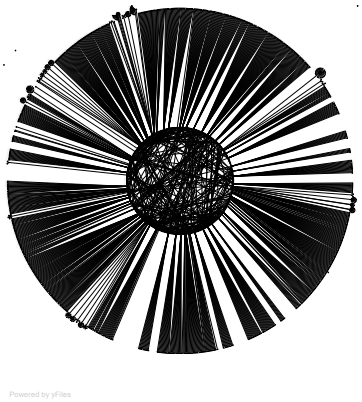

Figure 1: The BCC of target site U-1.

Figure 1 is a picture of a successful spammer U-1. The experiment revealed a neighborhood of 1380 sites, 266 of which were connected with 593 edges into a BCC forming the 3-trust neighborhood of the target site. The density and the size of the BCC is comparable to some of the larger BCCs of sites that openly promote "reciprocal linking".

The algorithm we outlined in the previous section would be very difficult to implement completely and efficiently at the browser side. We convert it to a heuristic that can be implemented on an average workstation and produce results within minutes. The goal of implementing a quick algorithm introduces a number of simplifications explained below.

To gather the back links of a site we use the Google API [15]. Still, some sites can have thousands of back links while others have only a few. For this experiment, we limited the number of backlinks for each site to 30. We call this parameter the backlink *fan*.

Determining the similarity of a site to the starting site was done by sampling a few pages of each site. We have noticed from other experiments that when one samples 5–10 pages per site one can get an excellent similarity measure. To speed things up in this experiment, however, we only sampled two pages per site. Even such a small sample, though, produced more than acceptable results.

Similarity was determined using the *df.idf* ranking on the universe of the sites explored. To decide on the cutoff point between LF's and MAS's, we created a site that contained a few random news pages from Reuters. We call this site the *divider* site. Pages more similar to the target site than is the divider site are categorized as LF sites; others are grouped into MAS sites.

Finally, to further reduce the effect of the explosive nature of the web, we introduced the concept of *stop sites*. A stop site is one that the user believes should not be included in the trust graph either because the trustworthiness of such a site is known or because it cannot be defined. In the first group we placed educational institutions as determined by their URL. In the latter we placed a few well known Directories and Blog sites.

We recognize that each of the decisions above can be strengthened, thus strengthening the results of our approach. In particular, increasing the fan parameter above 30 will recognize more sites in the neighborhood. Using a larger universe in calculating similarity and increasing the sampling of pages per site, will give a better approximation of the LF and MAS groups. Our rather conservative approach provided a solid proof of concept for our hypothesis, while remaining fast enough for browser-side implementation.

We ran a breadth-first search on the backlinks for each target site looking at the 3-trust neighborhood with a fan of 30 sites. We categorized the sites in the neighborhood into members of a link farm (LF) or members of a mutual admiration society (MAS) based on their similarity to the target site. The sites were then evaluated for trustworthiness. Due to the effort involved, only a randomly chosen 10% of the MAS sites were evaluated. All of the LF sites were evaluated, however.

As you can see from the results below, there were almost no trustworthy sites in the 3-trust neighborhoods of the untrustworthy ones. As one might expect, a trustworthy site is unlikely to deliberately link to an untrustworthy one, or even to one that "associates" with one. Not surprisingly, the statement is not as strong for the trustworthy sites, since untrustworthy sites are free to link to whomever they choose (although thanks to PageRank, spammers are unlikely to want to link to too many sites outside their spamming network in order to avoid "leaking" rank [5]).

Target	T(LF)	T(MAS)	U(LF)	U(MAS)
U-1	0%	0%	96%	88%
U-2	0%	0%	100%	65%
U-3	0%	0%	95%	100%
U-4	0%	0%	88%	83%
U-5	0%	0%	100%	73%
U-6	11%	9%	89%	57%
T-1	86%	70%	14%	0%
T-2	64%	64%	33%	13%

Our experiments showed that the quality of the starting site was a very good predictor for the quality of the BCC sites. While most of the results above show no accidental or erroneous linking of a trustworthy site to an untrustworthy one, we found such evidence in one experiment.

6. RELATED WORK

Web spamming has received a lot of attention lately [1, 3, 4, 5, 12, 13, 19, 21, 22, 24, 27, 29, 30, 33]. The first papers to raise the issue were [30, 22]. The spammers' success was noted in [13, 12, 10, 2, 4, 16, 23].

Characteristics of spamming sites based on diversion from power laws are presented in [12]. Current tricks employed by spammers are detailed in [18]. An analysis of the popular PageRank method employed by many search engines today and ways to maximize it in a spamming network is described in [5]. A modification to the PageRank to take into account opinions of human editors, employees of a search engine, is presented in [19]. A comprehensive treatment on social networks is presented in [38]. The connection between web spammers and social propagandists, and how the evolution of search engines can be understood as response to spammers is presented in [32]. Propagation methods for trust and distrust are discussed in [17]. Some work on personalized web search is presented in [20, 25]. The effect that search engines have on page popularity was discussed in [9].

7. CONCLUSIONS AND FUTURE DIRECTIONS

In this paper we have argued that web spam is to cyberworld what propaganda is to society. As far as we know, this is the first time this relationship is noted. As evidence of this analogy and its importance, we have shown that the evolution of search engines can be simply understood as the search engines' response to defend against spam. New search engines are not invented every few years, as it is sometimes reported; they are developed when researchers have a good answer to spam. More importantly, we have shown that this relationship can guide us towards developing heuristics that recognize spammers. In particular, we have presented automatic ways of recognizing trust neighborhoods on the web based on the biconnected component around some starting site. Experimental results from a number of such instances show our algorithm's ability of recognizing parts of a spamming network.

With such results, the question arises as to what one should do once one recognizes a spamming network. This is a question that has not attracted much attention in the past. The "obvious" approach is that a search engine would delete such networks from its indices [12] or might downgrade them by some prespecified amount [19] as it has been reported in the past [36].

Both of these approaches, however, require a universal agreement of what constitutes spam. Such an agreement cannot exist; one person's spam may be another person's treasure. Should the search engines determine what is trustworthy and what is not? Willing or not, they are the *de facto* arbiters of what information users see. As in a well-known cartoon, the kid responds to the old man who has been looking all his life for the meaning of life: "If it is not on Google or eBay, it does not exist."

We believe that it is the users' right and responsibility to decide what is acceptable for them. Their browser, their window to cyberworld, should enhance their ability to make this decision. User education is fundamental: People should know how search engines work and why, and how information appears on the web. But they should also have a browser that can help them determine the validity and trustworthiness of information.

The tool we described in an earlier section is a first step in this direction. Ultimately, it would be used along with a set of trust certificates that contains the portable trust preferences of the user, a set of preferences that the user can accumulate over time. A combination of search engines capable of providing indexed content and structure, including identified neighborhoods, with a browser capable of filtering those neighborhoods through the user's trust preferences,

would provide a new level of reliability to the user's information gathering.

8. ACKNOWLEDGEMENTS

The authors would like to thank Mirena Chausheva, Meredith Beaton-Lacoste and Scott Dynes for their valuable contributions.

9. REFERENCES

[1] B. Amento, L. Terveen, and W. Hill. Does authority mean quality? Predicting expert quality ratings of web documents. In *Proceedings of the Twenty-Third Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 2000.

[2] A. Arasu, J. Cho, H. Garcia-Molina, A. Paepcke, and S. Raghavan. Searching the web. *ACM Transactions on Internet Technology*, 1(1):2–43, June 2001.

[3] K. Bharat, A. Z. Broder, J. Dean, and M. R. Henzinger. A comparison of techniques to find mirrored hosts on the WWW. *Journal of the American Society of Information Science*, 51(12):1114–1122, 2000.

[4] K. Bharat, B.-W. Chang, M. R. Henzinger, and M. Ruhl. Who links to whom: Mining linkage between web sites. In *Proceedings of the 2001 IEEE International Conference on Data Mining*, pages 51–58. IEEE Computer Society, 2001.

[5] M. Bianchini, M. Gori, and F. Scarselli. PageRank and web communities. In *Web Intelligence Conference 2003*, Oct. 2003.

[6] S. Brin and L. Page. The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and ISDN Systems*, 30(1–7):107–117, 1998.

[7] A. Broder. A taxonomy of web search. *SIGIR Forum*, 36(2):3–10, 2002.

[8] A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, and J. Wiener. Graph structure in the web. In *Proceedings of the 9th international World Wide Web conference on Computer networks : the international journal of computer and telecommunications netowrking*, pages 309–320. North-Holland Publishing Co., 2000.

[9] J. Cho and S. Roy. Impact of search engines on page popularity. In *WWW 2004*, May 2004.

[10] T. S. Corey. Catching on-line traders in a web of lies: The perils of internet stock fraud. Ford Marrin Esposito, Witmeyer & Glesser, LLP, May 2001. http://www.fmew.com/archive/lies/.

[11] G. Cybenko, A. Giani, and P. Thompson. Cognitive hacking: A battle for the mind. *Computer*, 35(8):50–56, 2002.

[12] D. Fetterly, M. Manasse, and M. Najork. Spam, damn spam, and statistics. In *WebDB2004*, June 2004.

[13] D. Fetterly, M. Manasse, M. Najork, and J. Wiener. A large-scale study of the evolution of web pages. In *Proceedings of the twelfth international conference on World Wide Web*, pages 669–678. ACM Press, 2003.

[14] G. W. Flake, S. Lawrence, C. L. Giles, and F. Coetzee. Self-organization of the web and identification of communities. *IEEE Computer*, 35(3):66–71, 2002.

[15] I. Google. The Google api. http://www.google.com/apis/.

[16] L. Graham and P. T. Metaxas. "Of course it's true; i saw it on the internet!": Critical thinking in the internet era. *Commun. ACM*, 46(5):70–75, 2003.

[17] R. Guha, R. Kumar, P. Raghavan, and A. Tomkins. Propagation of trust and distrust. In *WWW 2004*, May 2004.

[18] Z. Gyongui and H. Garcia-Molina. Web spam taxonomy. Technical Report TR 2004-25, Stanford University, 2004.

[19] Z. Gyongui, H. Garcia-Molina, and J. Pedersen. Combating web spam with TrustRank. In *VLDB 2004*, Aug. 2004.

[20] T. H. Haveliwala. Topic-sensitive pagerank. In *Proceedings of the eleventh international conference on World Wide Web*, pages 517–526. ACM Press, 2002.

[21] M. R. Henzinger. Hyperlink analysis for the web. *IEEE Internet Computing*, 5(1):45–50, 2001.

[22] M. R. Henzinger, R. Motwani, and C. Silverstein. Challenges in web search engines. *SIGIR Forum*, 36(2):11–22, 2002.

[23] M. Hindman, K. Tsioutsiouliklis, and J. Johnson. Googlearchy: How a few heavily-linked sites dominate politics on the web. In *Annual Meeting of the Midwest Political Science Association*, April 3-6 2003.

[24] L. Introna and H. Nissenbaum. Defining the web: The politics of search engines. *Computer*, 33(1):54–62, 2000.

[25] G. Jeh and J. Widom. Scaling personalized web search. In *Proceedings of the twelfth international conference on World Wide Web*, pages 271–279. ACM Press, 2003.

[26] J. M. Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 46(5):604–632, 1999.

[27] R. Kumar, P. Raghavan, S. Rajagopalan, and A. Tomkins. Trawling the Web for emerging cyber-communities. *Computer Networks (Amsterdam, Netherlands: 1999)*, 31(11–16):1481–1493, 1999.

[28] A. Lee and E. L. (eds.). *The Fine Art of Propaganda*. The Institute for Propaganda Analysis. Harcourt, Brace and Co., 1939.

[29] C. A. Lynch. When documents deceive: trust and provenance as new factors for information retrieval in a tangled web. *J. Am. Soc. Inf. Sci. Technol.*, 52(1):12–17, 2001.

[30] M. Marchiori. The quest for correct information on the web: hyper search engines. *Comput. Netw. ISDN Syst.*, 29(8-13):1225–1235, 1997.

[31] M. L. Maulding. Lycos: Design choices in an internet search service. *IEEE Expert*, January-February( ):8–11, 1997.

[32] P. T. Metaxas. Web spam: An application of Propaganda theory. Technical Report CSD-TR27-2004, Wellesley College, 2004.

[33] G. Pringle, L. Allison, and D. L. Dowe. What is a tall poppy among web pages? In *Proceedings of the seventh international conference on World Wide Web 7*, pages 369–377. Elsevier Science Publishers B. V., 1998.

[34] G. Salton. Dynamic document processing. *Commun. ACM*, 15(7):658–668, 1972.

[35] C. Silverstein, H. Marais, M. Henzinger, and M. Moricz. Analysis of a very large web search engine query log. *SIGIR Forum*, 33(1):6–12, 1999.

[36] M. Totty and M. Mangalindan. As google becomes web's gatekeeper, sites fight to get in. In *Wall Street Journal CCXLI(39)*, February 26 2003.

[37] A. Vedder. Medical data, new information technologies and the need for normative principles other than privacy rules. In *Law and Medicine. M. Freeman and A. Lewis (Eds.), (Series Current Legal Issues)*, pages 441–459. Oxford University Press, 2000.

[38] S. Wasserman and K. Faust. *Social Network Analysis: Methods and Applications*. Cambridge University Press, 1994.