Enhancing Information Reliability through Backwards Propagation Of Distrust

Panagiotis Metaxas
Computer Science Department
Wellesley College
106 Central Street, Wellesley, MA 02481, USA
Email: pmetaxas@wellesley.edu

***Abstract*—Search Engines have greatly influenced the way we experience the web. Since the early days of the web people have been relying on search engines to find useful information. However, their ability to provide useful and unbiased information can be manipulated by Web spammers. 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. But it is mainly a serious problem for web users because they tend to confuse trusting the search engine with trusting the results of a search.**

In this paper, first we discuss the relationship between Web spam in cyber world and social propaganda in the real world. Then, we propose "backwards propagation of distrust," as an approach to finding spamming untrustworthy sites. Our approach is inspired by the social behavior associated with distrust. In society, recognition of an untrustworthy entity (person, institution, idea, etc) is a reason for questioning the trustworthiness of those that recommended this entity. People that are found to strongly support untrustworthy entities become untrustworthy themselves. In other words, in the society, distrust is propagated backwards.

Our algorithm simulates this social behavior on the web graph with considerable success. Moreover, by respecting the user's perception of trust through the web graph, our algorithm makes it possible to resolve the moral question of who should be making the decision of weeding out untrustworthy spammers in favor of the user, not the search engine or some higher authority. Our approach can lead to browser-level, or personalized server-side, web filters that work in synergy with the powerful search engines to deliver personalized, trusted web results.

An earlier version of this paper was presented at [35].

***Keywords*-Web search; Information Reliability; Web graph; Link structure; Web Trust; Web Spam**

I. Introduction

Search Engines have greatly influenced the way we experience the web. Since the early days of the web people have been relying on search engines to find useful information. When the web was relatively small, Web directories were built and maintained that were using human experts to screen and categorize pages according to their characteristics. By the mid 1990's, however, it was apparent that the human expert model of categorizing web pages would not scale. The first search engines appeared and they have been evolving ever since.

But what influences their evolution? The way a user interacts with a search engine is through the search results to a query that he or she has issued. Search engines know that the quality of their ranking will determine how successful they are. If users perceive the results as valuable and reliable, they will come again. Otherwise, it is easy for them to switch to another search engine.

Research in Information Retrieval has produced a large body of work that, theoretically, produces high quality search results. Yet, search engines admit that IR theory is but one of their considerations. One of the major issues that influences the quality of ranking is the effect that web spam has on their results. *Web spamming* is defined as the practice of manipulating web pages in order to influence search engines rankings in ways beneficial to the spammers. Spammers aim at search engines, but target the end users. Their motive is usually commercial, but can also be political or religious.

We should mention here that, to people unfamiliar with web spam, the term is confused with email spam. Even though both term describe manipulation of information to confuse people in cyberspace, which is why we call them both "spam", they are very different in the way we experience them. In particular, email spam is pushed onto the users through email and we can learn to recognize it easily. Web spam, on the other hand, is misinformation that we pull through search engines, and thus it is very difficult to learn to recognize it. Sometimes, the term "adversarial information retrieval" is used to describe web spam. A more descriptive name for it would be "search engine ranking manipulation."

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 exploit this trust.

From their side, the search engines have put considerable effort in delivering spam-free query results and have developed sophisticated ranking strategies. Two such ranking strategies that have received major attention are PageRank

[5] and HITS [27]. 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[21]. Search engines typically see web spam as an interference to their operations and would like to restrict it, but there can be no algorithm that can recognize spamming sites based solely on graph isomorphism [4].

In this paper, we first explain why we need to understand web spamming beyond the technical details. Web spamming is a social problem first, then a technical one, and there is strong relationship between it and social propaganda. In fact, [34] presents evidence of its influence on the evolution of search engines. Then, we describe and evaluate an algorithmic way of discovering spamming networks automatically. Finally, we discuss a general framework for the long-term approach to web spam

A. Background

Web spamming has received a lot of attention lately [3], [4], [12], [13], [17], [20], [21], [23], [28], [31], [32], [37], [34]. The first papers to raise the issue were [32], [21]. The spammers' success was noted in [3], [10], [12], [13], [22]. Web search was explained in [1]. The related topic of cognitive hacking was introduced in [11].

Characteristics of spamming sites based on diversion from power laws are presented in [12]. Current tricks employed by spammers are detailed in [16]. 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 [4]. TrustRank, a modification to the PageRank to take into account the evaluations of a few seed pages by human editors, employees of a search engine, is presented in [17]. Techniques for identifying automatically link farms of spam pages were presented in [45], [2].

A comprehensive treatment on social networks is presented in [43]. The connection between the Web and social networks was explicitly noted in [29], [38] and implicitly used in [5], [27]. In fact, Kleinberg's work explores many of these connections (e.g., [26]). Identification of web communities was explored in [28], [14]. Work on topic-sensitive and personalized web search is presented in [19], [24]. The effect that search engines have on page popularity was discussed in [8].

Research in the past has focused on the identification of web communities through the use of bipartite cores [28] or maximum flow in dense subgraphs [14]. Some of the background information on Web Spam and its connection to social propaganda was presented in [34].

The rest of this paper is organized as follows. The next section gives an overview of the problem of information reliability and web spamming. Section II-B has a short introduction to the theory of propaganda detection and the next section II-C discusses the relationship between the Web Graph and the trust social network. The following section II-D analyzes the evolution of search engines as their response to spam. Section III describes the backward propagation of distrust algorithm and the following section IV presents some of our experimental results running this algorithm. Finally, the last section V has our conclusions and a framework for the long-term approach to web spam.

II. On Information Reliability and Web Spam

A. Web Spam

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 on 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 in their lives [10], [42], [22], [31]. Moreover, 85% of the time, people do not look past the first ten results returned by the search engine [40]. 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 "Search Engine Optimization" 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 [12].

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:

Text manipulations: This includes repeating text excessively 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 manipulations: This technique aims to change the perceived structure of the Web Graph in order to cause incorrect calculation of page reputation. Such examples are the so-called "link-farms," domain flooding (plethora of domains that re-direct to a target site), page "awards," (the spammer pretends to run an organization that distributes awards for web site design or information; the awarded site gets to display the "award", an image linking back to awarding organization, effectively increasing the visibility of the spammer' site), etc.

Both kinds of spam aim to boost the ranking of spammed web pages. So as not to get caught, spammers conceal their

actions through cloaking, content hiding and redirection. Cloaking, for example, aims to serve different pages to search engine robots and to web browsers (users). The spamming 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, 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 not spam, but a tool that helps spammers hide their attacks. For a comprehensive treatment of the spamming techniques, see [16].

Since anyone can be an author on the web, these practices have brought into prominence 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. A recent study [15] 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." Very few of them would double-check for validity. 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] puts the blame squarely on the questionable critical thinking skills of the investors for the success of stock fraud cases.

B. Social Theory of Propaganda

On the outset, it may seem surprising that a technical article discusses social propaganda. This is a subject that has been studied extensively by social scientists and might seem out of the realm of computing. However, the web is a social network, influenced daily by the actions (intentional or otherwise) of millions of people. In that respect, web researchers should be aware of social theories and practices since they may have applicability in their work. We believe that a basic understanding of social propaganda can be valuable to technical people designing and using systems that affect our social interactions. In particular, it can be useful to researchers that study Web Spam. We offer here a brief introduction to the theory of propaganda detection.

There are many definitions of propaganda, reflecting its multiple uses over time. One working definition we will use here is

Propaganda is the attempt to modify human behavior, and thus influence people's actions in ways beneficial to propagandists.

Propaganda has a long history in modern society and is often associated with negative connotation. This was not always the case, however. The term was first used in 1622, in the establishment by the Catholic Church of a permanent Sacred Congregation *de Propaganda Fide* (for the propagation of faith), a department which was trying to spread Catholicism in non-Catholic Countries [44]. Its current meaning comes from the successful Enemy Propaganda Department in the British Ministry of Information during WWI. However, it was not until 1938, in the beginning of WWII, that a theory was developed to detect propagandistic techniques. For the purposes of this paper we are interested in ways of detecting propaganda, especially by automatic means.

First developed by the Institute for Propaganda Analysis [30], classic Propaganda Theory identifies several techniques that propagandists often employ in order to manipulate perception.

- **Name Calling** is the practice of giving an idea a bad label. It is used to make people reject and condemn the idea without examining the evidence. For example, using the term "miserable failure" to refer to political leaders such as US President George Bush can be thought of as an application of name calling.
- **Glittering Generalities** is the mirror image[1] of name calling: Associating an idea with a "virtue word", in an effort to make us accept and approve the idea without examining the evidence. For example, using the term "patriotic" to refer to illegal actions is a common application of this technique.
- **Transfer** is the technique by which the propagandist carries over the authority, sanction, and prestige of something respected and revered to something he would have us accept. For example, delivering a political speech in a mosque or a church, or ending a political gathering with a prayer have the effect of transfer.
- **Testimonial** is the technique of having some respected person comment on the quality of an issue on which they have no qualifications to comment. For example, a famous actor who plays a medical doctor on a popular TV show tells the viewers that she only uses a particular pain relief medicine. The implicit message is that if a famous personality trusts the medicine, we should too.
- **Plain Folks** is a technique by which speakers attempt to convince their audience that they, and their ideas, are "of the people," the "plain folks". For example, politicians sometimes are seen flipping burgers at a neighborhood diner.
- **Card Stacking** involves the selection of facts (or falsehoods), illustrations (or distractions), and logical

[1]Name calling and glittering generalities are sometimes referred to as "word games."

(or illogical) statements in order to give an incorrect impression. For example, some activists refer to the Evolution Theory as a theory teaching that humans came from apes (and not that both apes and humans have evolved from a common ancestor who was neither human nor ape).

- **Bandwagon** is the technique with which the propagandist attempts to convince us that all members of a group we belong to accept his ideas and so we should "jump on the band wagon". Often, fear is used to reinforce the message. For example, commercials might show shoppers running to line up in front of a store before it is open.

The reader should not have much trouble identifying additional examples of such techniques used in politics or advertising. The next section discusses the relationship of propaganda to web spam, by first describing the similarity of social networks to the web graph.

C. The Web Graph as a Trust Network

The web is typically represented by a directed graph [7]. The nodes in the Web Graph are the pages (or sites) that reside on servers on the internet. Arcs correspond to hyperlinks that appear on web pages (or sites). In this context, web spammers' actions can be seen as altering the contents of the web nodes (mainly through text spam), and the hyperlinks between nodes (mainly through link spam).

The theory of social networks [43] also uses directed graphs to represent relationships between social entities. The nodes correspond to social entities (people, institutions, ideas). Arcs correspond to recommendations between the entities they connect. In this context, propagandistic techniques can be seen as altering the trust social network by altering one or more of its components (i.e., nodes, arcs, weights, topology).

To see the correspondence more clearly, we will examine some of the propagandistic techniques that have been used successfully by spammers: The technique of testimonials effectively adds a link between previously unrelated nodes. Glittering generalities change the contents of a node, effectively changing its perceived relevance. Mislabeled anchor text is an example of card stacking. And the technique of bandwagon creates many links between a group of nodes, a "link farm". So, we define web spam based on the spammers actions:

Web Spam is the attempt to modify the web (its structure and contents), and thus influence search engine results in ways beneficial to web spammers.

Table I has the correspondence, in graph theoretic terms, between the web graph according to a search engine and the trust social network of a particular person. Web pages or sites correspond to social entities and hyperlinks correspond to trust opinions. The rank that a search engine assigns to a page or a site corresponds to the reputation a social entity has for the person. This rank is based on some ranking formula that a search engine is computing, while the reputation is based on idiosyncratic components associated with the person's past experiences and selective application of critical thinking skills; both are secret and changing.

This correspondence is more than a coincidence. The web itself is a social creation, and both PageRank and HITS are socially inspired ranking formulas. [5], [27], [38], [1]. Socially inspired systems are subject to socially inspired attacks. Not surprisingly then, the theory of propaganda detection can provide intuition into the dynamics of the web graph.

PageRank is based on the assumption that the reputation of an entity (a web page in this case) can be measured as a function of both the number and reputation of other entities linking to 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[2]. The implicit assumption is that hyperlink "voting" is taking place independently, without prior agreement or central control. Spammers, like social propagandists, form structures that are able to gather a large number of such "votes of confidence" by design, thus breaking the crucial assumption of independence in a hyperlink. But while the weights in the web graph are assigned by each search engine, the weights in the trust social network are assigned by each person. Since there are many more persons than search engines, the task of a web spammer is far easier than the task of a propagandist.

D. Search Engine Evolution

In the early 90's, when the web numbered just a few million servers, the **first generation** search engines were ranking search results using the vector model ([39], [20]) of classic information retrieval techniques: the more rare words two documents share, the more similar they are considered to be.

According to the *vector model* in Information Retrieval [39], documents contained in a document collection D are viewed as vectors in term space T. Under this formulation, rare words have greater weight than common words, because they are viewed as better representing the document contents. In the vector model, document similarity $sim(D_1, D_2)$ between document vectors D_1 and D_2 is represented by the angle between them. A search query Q is considered simply a short document and the results of a search for Q are ranked according to their (normalized) similarity to the query. While the exact details of the computation of term weights were kept secret, we can say that the ranking formula R^{G_1} in the first generation search engines was based in the following

[2]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. Recently, Google introduced the "nofollow" attribute for hyperlinks, as a tool for blog site owners to mark visitor opinions. It is very unlikely that spamming blog owners will use it, however.

Graph Theory	*Web Graph*	*Trust Social Network*
Node	web page or site	social entity
weight	rank (accord. to a search engine)	reputation (accord. to a person)
weight computation	ranking formula (e.g., pagerank)	idiosyncratic (e.g., 2 recommenders)
	computed continuously	computed on demand
Arc	hyperlink	trust opinion
semantics	"vote of confidence"	"recommendation"
weight	degree of confidence	degree of entrustment
weight range	$[0 \dots 1]$	$[distrust \dots trust]$

Table I
GRAPH THEORETIC CORRESPONDENCE BETWEEN THE WEB GRAPH AND THE TRUST SOCIAL NETWORK. THERE IS A ONE-TO-ONE CORRESPONDENCE BETWEEN EACH COMPONENT OF THE TWO GRAPHS. A MAJOR DIFFERENCE, HOWEVER, IS THAT, EVEN THOUGH A PERSON MAY FEEL NEGATIVE TRUST (DISTRUST) FOR SOME ENTITY, THERE IS NO NEGATIVE WEIGHT FOR HYPERLINKS.

principle: the more rare keywords a document shares with a query, the higher similarity it has with it, resulting in a higher ranking score for this document:

$$R^{G_1} = f(sim(p, Q)) \quad (1)$$

The first attack to this ranking came from within the search engines. In 1996, search engines started openly selling search keywords to advertisers [9] 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.

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 using **keyword stuffing**, i.e., 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 [30, pg. 47]:

The propagandist associates one or more suggestive words without evidence to alter the conceived value of a person or idea.

In an effort to nullify the effects of glittering generalities, **second generation** search engines started employing additionally more sophisticated ranking techniques. One of the more successful techniques was based on the "link voting principle": Each web site s has value equal to its "popularity" $|B_s|$ which is influenced by the set B_s of sites pointing to s.

Therefore, the more sites were linking to a site s, the higher the popularity of s's pages. Lycos became the champion of this ranking technique [33] and had its own popularity skyrocket in late 1996. Doing so, it was also distancing itself from the ethical questions introduced by blurring advertising with ranking [9].

The ranking formula R^{G_2} in the second generation search engines was a combination of a page's similarity, $sim(p, Q)$, and its site's popularity $|B_s|$:

$$R^{G_2} = f(sim(p, Q), |B_s|) \quad (2)$$

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 were figured out by experimentation and reverse engineering. (e.g., [37], [32]).

Unfortunately, this ranking formula 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, a technique that subsequently became known as **link farms**. The link voting principle was socially inspired, so spammers used the well known propagandistic method of *bandwagon* to circumvent it [30, pg. 105]:

With it, the propagandist attempts to convince us that all members of a group to which we belong are accepting his program and that we must therefore follow our crowd and "jump on the band wagon".

Similarly, the spammer is promoting the impression of a high degree of popularity by inter-linking many internally controlled sites that will eventually all share high ranking.

PageRank and HITS marked the development of the **third generation** search engines. The introduction of PageRank in 1998 [5] was a major event for search engines, because it seemed to provide a more sophisticated anti-spamming solution. Under PageRank, not every link contributes equally to the "reputation" of a page $PR(p)$. Instead, links from highly reputable pages contribute much higher value than links from other sites. That way, the link farms developed by spammers would not influence much their PageRank, and Google became the search engine of choice. HITS is another socially-inspired ranking which has also received a lot of attention [27] and is reportedly used by the AskJeeves search engine. 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 and they both gain reputation.

Unfortunately, spammers again found ways of circumventing these rankings. In PageRank, a page enjoys absolute reputation: its reputation is not restricted on some particular issue. Spammers deploy sites with expertise on irrelevant

subjects, and they acquire (justifiably) high ranking on their expert sites. Then they bandwagon the irrelevant expert sites, creating what we call a **mutual admiration society**. In propagandistic terms, this is the technique of *testimonials* [30, pg. 74] often used by advertisers:

Well known people (entertainers, public figures, etc.) offer their opinion on issues about which they are not experts.

Spammers were so aggressive in pursuing this technique that they openly promoted "reciprocal links": Web masters controlling sites that had some minimum PageRank, were invited to join a mutual admiration society by exchanging links, so that at the end everyone's PageRank would increase. HITS has also shown to be highly spammable by this technique due to the fact that its effectiveness depends on the accuracy of the initial neighborhood calculation.

Another heuristic that third generation search engines used was that of exploiting "anchor text". It had been observed that users creating links to web pages would come to use, in general, meaningful descriptions of the contents of a page. (Initially, the anchor text was non-descriptive, such as "click here", but this changed in the late 1990's.) Google was the first engine to exploit this fact noting that, even though IBM's web page made no mention that IBM is a computer company, many users linked to it with anchor text such as "computer manufacturer".

Spammers were quick to exploit this feature too. In early 2001, a group of activists started using the anchor text "miserable failure" to link to the official Whitehouse page of American President George W. Bush. Using what became known as "Googlebomb" or, more accurately, **link-bomb** since it does not pertain to Google only, other activists linked the same anchor text to President Carter, filmmaker Michael Moore and Senator Hilary Clinton.

Using the anchor text is socially inspired, so spammers used the propagandistic method of *card stacking* to circumvent it [30, pg. 95]:

Card stacking involves the selection and use of facts or falsehoods, illustrations or distractions, and logical or illogical statements in order to give the best or the worst possible case for an idea, program, person or product.

The ranking formula R^{G_3} in the third generation search engines is, therefore, some secret combination of a number of features, primarily the page's similarity, $sim(p, Q)$, its site's popularity $|B_s|$ and its the page's reputation $PR(p)$:

$$R^{G_3} = f(sim(p, Q), |B_s|, PR(p)) \quad (3)$$

Search engines these days claim to have developed hundreds of little heuristics for improving their web search results [18] but no big idea that would move their rankings beyond the grasp of spammers. As Table II summarizes, for every idea that search engines have used to improve their ranking, spammers have managed quickly to balance it with techniques that resemble propagandistic techniques from society. Web search corporations are reportedly busy developing the engines of the next generation [6]. The new techniques aim 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?

This may not be possible. Our analysis shows that we are trying to create in cyberspace what societies have not succeeded in creating in their real space. However, we can learn to live in a web with spam as we live in society with propaganda, given appropriate education and technology.

III. An Anti-propagandistic Algorithm

Since spammers employ propagandistic techniques [34], it makes sense to design anti-propagandistic methods for defending against them. These methods need to be user-initiated, that is, the user decides which web site not to trust and then seeks to distrust those supporting the untrustworthy web site. 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 design methods that help individuals make more informed decisions about the quality of the information they find on the web.

Here is one way that people defend against propaganda in every day life:

In society, distrust is propagated backwards: When an untrustworthy recommendation is detected, it gives us a reason to reconsider the trustworthiness of the recommender. Recommenders who strongly support an untrustworthy recommendation become untrustworthy themselves.

This process is selectively repeated a few times, propagating the distrust backwards to those who strongly support the recommendation. The results of this process become part of our belief system and are used to filter future information. (Note that distrust is not propagated forward: An untrustworthy person's recommendations could be towards *any* entity, either trustworthy or untrustworthy.)

We set out to test whether a similar process might work on the web. Our algorithm takes as input s, a web site, which is represented by the URL of the server containing a page that the user determined to be untrustworthy. This page could have come to the user through web search results, an email spam, or via the suggestion of some trusted associate (e.g., a society that the user belongs to).

The obvious challenge in testing this hypothesis would be to retrieve a neighborhood of web sites linking to the starting site s in order to analyze it. Since we are interested in back links to sites, we can not just follow a few forward links (hyperlinks on web sites) to get this information. Otherwise we would need to possibly explore the whole web graph. Today, only search engines have this ability. Thankfully, search engines have provided APIs to help with our task.

S.E.'s	Ranking	Spamming	Propaganda
1st Gen	Doc Similarity	keyword stuffing	glittering generalities
2nd Gen	+ Site popularity	+ link farms	+ bandwagon
3rd Gen	+ Page reputation + anchor text	+ mutual admiration societies + link bombs	+ testimonials + card stacking

Table II
CHANGES IN RANKING BY GENERATIONS OF SEARCH ENGINES, THE RESPONSE OF THE WEB SPAMMERS AND THE CORRESPONDING PROPAGANDISTIC TECHNIQUES.

Starting from s we build a breadth-first search (BFS) tree of the sites that link to s within a few "clicks" (Figure 1). We call the directed graph that is revealed by the back-links, the "trust neighborhood" of s. We do not explore the web neighborhood directly in this step. Instead, we can use the Google API for retrieving the back-links.

Referring to Figure 1, if one deems that starting site 1 is untrustworthy, and sites 2, 3, 4, 5 and 6 link directly to it, one has reasons to be suspicious of those sites too. We can take the argument further and examine the trustworthiness of those sites pointing to 2, ... 6. The question arises on whether we should distrust all of the sites in the trust neighborhood of starting site s or not. Is it reasonable to become suspicious of every site linking to s in a few steps? They are "voting in confidence" after all [5], [27]. Should they be penalized for that? Such a radical approach is not what we do in everyday life. Rather, we selectively propagate distrust backwards only to those that most strongly support an untrustworthy recommendation. Thus, we decided to take a conservative approach and examine only those sites that use link spamming techniques in supporting s. In particular, we focused on the biconnected component (BCC) that includes s (Figure 2).

A BCC is a graph that cannot be broken into disconnected pieces by deleting any single vertex. An important characteristic of the BCC is there are at least two independent paths from any of its vertices to s. Strictly speaking, the BCC is computed on the undirected graph of the trust neighborhood. But since the trust neighborhood is generated through the BFS, the cross edges (in BFS terminology) create cycles in the undirected graph (Figure 1). Each cycle found in the BCC must have at least one "ring leader", from which there are two directed paths to s, one leaving through the discovery edge and the other through the cross edge. We view the existence of multiple paths from ring leaders to s as evidence of strong support of s. The BCC reveals the members of this support group. The graph induced by the nodes not in the BCC is called "BFS periphery".

More formally, the algorithm is as follows:

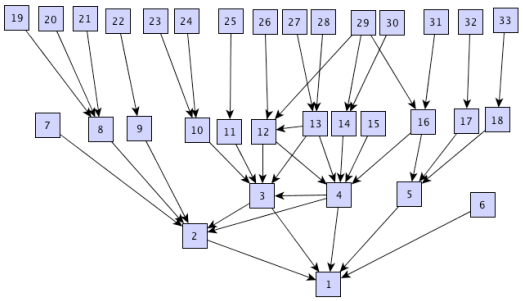

Figure 1. An example of a breadth-first search tree in the trust neighborhood of site 1. Note that some nodes (12, 13, 16 and 29) have multiple paths to site 1. We call these nodes "ring leaders" that show a concerted effort to support site 1.

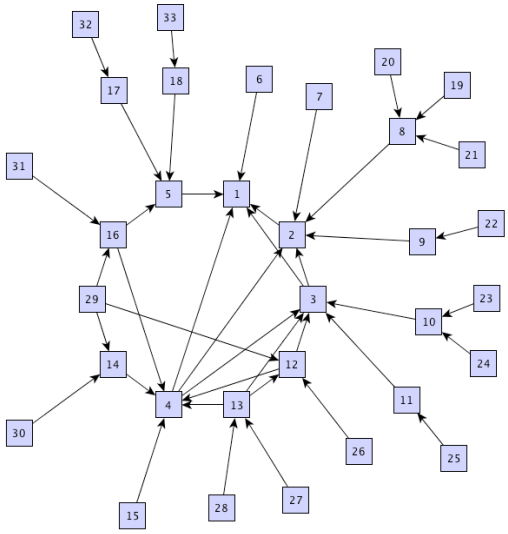

Figure 2. The BCC of the trust neighborhood of site 1 is drawn in a circular fashion for clarity. Note that the BCC contains the "ring leaders," that is, those nodes with multiple paths leading to s. The graph induced by the nodes not in the BCC is called "BFS periphery".

```
Input:
  s = Untrustworthy starting site's URL
  D = Depth of search
  B = Number of back-links to record

S = {s}
Using BFS for depth D do:
  Compute U={sites linking to sites in S}
    using the Google API
    (up to B back-links / site)
  Ignore blogs, directories, edu's
```

```
  S = S + U
Compute the BCC of S that includes s

Output: The BCC
```

A. Implementation Details

To be able to implement the above algorithm at the browser side, we restrict the following parameters: First, the BFS's depth D is set to 3. We are not interested in exploring a large chunk of the web, just a small neighborhood around s. Second, we limit the number B of back-link requests from the Google API to 30 per site. This helps reduce the running time of our algorithm since the most time-consuming step is the query to Google's back-link database. Finally, we introduced in advance a set of "stop sites" that are not to be explored further.

A *stop site* is one that should not be included in the trust neighborhood either because the trustworthiness of such a site is irrelevant, or because it cannot be defined. In the first category we placed URLs of educational institutions (domains ending in .edu). Academicians are not in the business of linking to commercial sites [36]. When they do, they do not often convey trust in the site. College libraries and academicians, for example, sometimes point to untrustworthy sites as examples to help students critically think about information on the web. In the latter category we placed a few well known Directories (URLs ending in yahoo.com, dmoz.org, etc.) and Blog sites (URLs containing the string 'blog' or 'forum'). While blogs may be set up by well meaning people who are trying to increase the discourse on the web, blog pages are populated with opinions of many people and are not meant to represent the opinion of the owner. Anyone can put an entry into an unsupervised blog or directory, and following a hyperlink from a blog page should not convey the trustworthiness of the whole blog site. If the search engines were able to distinguish and ignore links inside the comments, blogs could be removed from the stop sites. No effort to create an exhaustive list of blogs or directories was made.

With these restrictions, our algorithm can be implemented on an average workstation and produce graphs with up to a few hundred nodes within minutes. As we mentioned, the most time demanding step is requesting and receiving the back-link lists from Google, since it requires initiating an online connection. No connections to the particular web sites was done during the creation of the trust neighborhood. Performing the BFS and computing the BCC of the graph assembled is done in time linear on the number of sites retrieved, so it is fast. We used the JUNG software library [25] to maintain the web subgraph and compute its BCC. The whole neighborhood can fit into the main memory of the workstation, so this does not require additional time.

IV. Finding Untrustworthy Neighborhoods that use Link Spam

There are several ways one can run into an initial untrustworthy site to use it as a starting site s. For example, search results for queries that happen to be controversial (e.g., "Armenian genocide", "morality of abortion" and "ADHD real disease") or happen to be the source of unreliable advertisement (e.g., "human growth hormone increase muscle mass"), contain plethora of responses that can be considered untrustworthy. In our experiments, we examined the trust neighborhoods of eight untrustworthy and two trustworthy sites. In Table III below these sites are labeled as U-1 to U-8 and T-1 to T-2, respectively. See Figure 3 for an example of U-1.

We run the experiments between September 17 and November 5, 2004. At the time of the experiment, all sites happen to have comparable PageRank, as reported by the Google Toolbar. In fact, U-1 and T-1 both had PageRank 6 while the remaining sites had PageRank 5. We recorded the PageRank numbers as reported by the Google Toolbar because this is always one of the first questions people ask and because the spamming industry seems to use it as a measure of their success. In fact, one can find spam networks inviting the creation of "reciprocal links" for sites that have at lease a minimum of PageRank 5, in order to increase their overall PageRank.

To determine the trustworthiness of each site we had a human evaluator look at a sample of the sites of the BCC. The results of our experiments appear on Table III. Due to the significant manual labor involved, only 20% of the total 1,396 BCC sites were sampled and evaluated. To select the sample sites, we employed stratified sampling with skip interval 5. The stratum used was similarity of the site to the starting site.

Each site in the sample was classified as either Trustworthy, Untrustworthy, or Non-determined. The last category includes a variety of sites for which the evaluator could not clearly classify.

We have two main results:

1. The trustworthiness of the starting site is a very good predictor for the trustworthiness of the BCC sites.

In fact (see Table 1), there were very few trustworthy sites in the trust neighborhoods of sites U-1 to U-8. The reason is, we believe, that a trustworthy site is unlikely (though not impossible) to deliberately link to an untrustworthy site, or even to a site that associates itself with an untrustworthy one. In other words, *the "vote of confidence" link analogy holds true only for sites that are choosing their links responsibly.* The analogy is not as strong when starting from a trustworthy site, since untrustworthy sites are free to link to whomever they choose. After all, there is some value in portraying a site in good company: Non-critically

Figure 3. The trust graph of starting site U-1. The circularly drawn nodes in the middle form its largest biconnected component. This experiment found a trust graph of 1307 sites, 228 of which were connected with 465 edges into a **bi-connected component (BCC)**. The central, circularly drawn component is the BCC, while the sites drawn on the **BCC Periphery** were the remaining 1079 sites discovered by the BFS algorithm. Only 2% trustworthy sites were found in the BCC, while 74% of them were untrustworthy. In contrast, 31% trustworthy and 33% untrustworthy sites were found in the BFS periphery. The remaining sites were mostly directories or other non-determined sites.

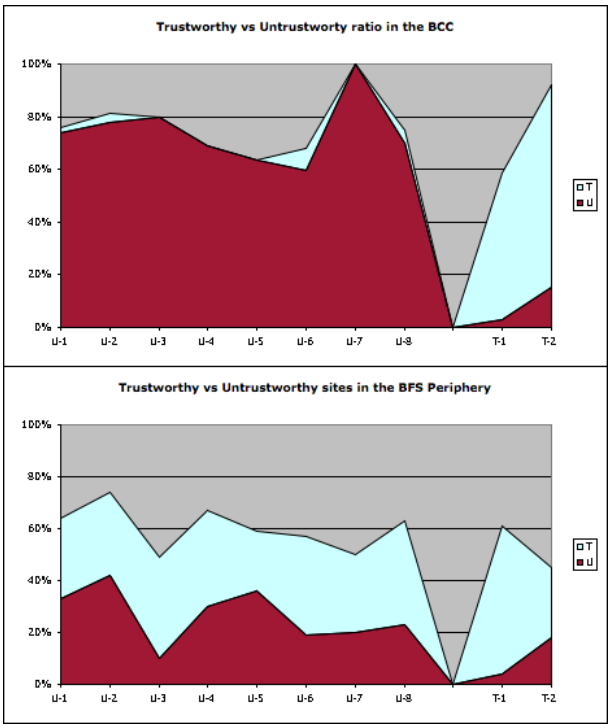

Figure 4. The trustworthy and untrustworthy percentages for trust neighborhoods of the BCC (top) and BFS peripheral (bottom) sites for the data shown in Table III. On the horizontal coordinates are shown 8 untrustworthy (on the left) and 2 trustworthy sites (on the right side of each graph). The vertical coordinates are the percentages of untrustworthy (U) and trustworthy (T) sites found in the neighborhood of each starting site. Comparing the left and right sides of the top graph, one can see that the trustworthiness of the starting site is a very good predictor for the trustworthiness of the BCC sites. Comparing the top and bottom graphs, one can see that the BCC is significantly more predictive of untrustworthy sites than the BFS periphery

thinking users may be tempted to conclude that, if a site points to "good" sites, it must be "good" itself.

2. THE BCC IS SIGNIFICANTLY MORE PREDICTIVE OF UNTRUSTWORTHY SITES THAN THE BFS PERIPHERY.

In particular (see Figure 4, top), in the BCC of an untrustworthy starting site, we found that, on average, 74% of the sites were also untrustworthy, while only 9% were trustworthy. In the BFS periphery (see Figure 4, bottom), these average percentages change to 27% untrustworthy and 11% trustworthy, with the rest non-determined. This suggests that the trustworthiness of sites in the BFS periphery is essentially unrelated to the trustworthiness of the starting site.

A. Future Directions: Incorporating Content Analysis

In our experiments we also devised a simple method to evaluate the similarity of the contents of each site to the starting site s. After the trust neighborhood was explored, we fetched and concatenated a few pages from each site (randomly choosing from the links that appeared in the domain URL) into a document. Then, we tried to determine the similarity of each such document to the document of the starting site. Similarity was determined using the $tf.idf$ ranking on the universe of the sites explored. We are aware that having a limited universe of documents does not give the best similarity results, but we wanted to get a feeling of whether our method could further be used to distinguish between "link farms" (spamming sites controlled by a single entity) and "mutual admiration societies" (groups of independent spammers choosing to exchange links). The initial results are encouraging, (see Fig. 5) showing a higher percentage of untrustworthy sites among those most similar to the starting site s.

Several possible extensions can be considered in this work. Generating graphs with more back-links per site, studying the evolution of trust neighborhoods over time, examining the density of the BCCs, and finding a more reliable way to compute similarity are some of them. We

S	$\|V_G\|$	$\|E_G\|$	$\|V_{BCC}\|$	$\|E_{BCC}\|$	**Trust**$_{BCC}$	**Untr**$_{BCC}$	**Trust**$_{BFS}$	**Untr**$_{BFS}$
U-1	1307	1544	228	465	2%	74%	31%	33%
U-2	1380	1716	266	593	4%	78%	32%	42%
U-3	875	985	97	189	0%	80%	39%	10%
U-4	457	509	63	115	0%	69%	37%	30%
U-5	716	807	105	189	0%	64%	23%	36%
U-6	312	850	228	763	9%	60%	38%	19%
U-7	81	191	32	143	0%	100%	30%	20%
U-8	1547	1849	200	430	5%	70%	40%	23%
T-1	1429	1566	164	273	56%	3%	57%	4%
T-2	241	247	13	17	77%	15%	27%	18%

Table III
SIZES OF THE EXPLORED TRUST NEIGHBORHOODS G AND THEIR BCC'S FOR EIGHT UNTRUSTWORTHY (U-1 TO U-8) AND TWO TRUSTWORTHY (T-1 AND T-2) STARTING SITES. $|V_G|$ CONTAINS THE NUMBER OF VERTICES AND $|E_G|$ THE NUMBER OF EDGES THAT OUR ALGORITHM FOUND IN THE TRUST NEIGHBORHOOD OF STARTING SITE s (STARTING FROM SITE s AND EXPLORING IN BFS MODE THEIR BACK-LINKS.) COLUMNS $|V_{BCC}|$ AND $|E_{BCC}|$ CONTAINS THE NUMBERS OF VERTICES AND EDGES OF THE LARGEST BICONNECTED COMPONENT WITHIN G. THE NEXT FOUR COLUMNS CONTAINS THE ESTIMATED PERCENTAGES OF TRUSTWORTHY AND UNTRUSTWORTHY SITES FOUND IN THE BCCS AND THE BFS PERIPHERIES (RESPECTIVELY). 20% OF EACH BCC AND 10% OF EACH BFS PERIPHERY WERE EVALUATED USING STRATIFIED SAMPLING.

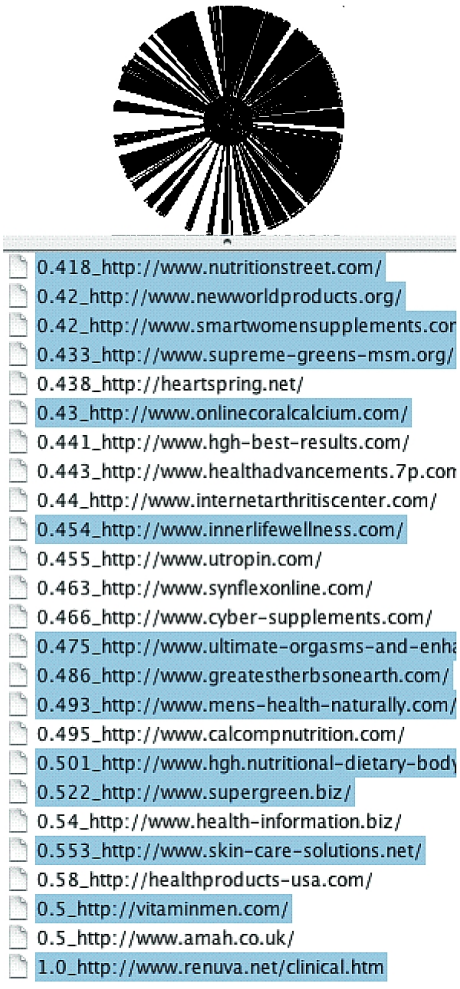

Figure 5. The list of sites similar to the starting site U-1 (at the end of the list). The highlighted sites are those that participate in the BCC. The decimal number in front of the URL corresponds to its calculated content similarity to the starting site (which has similarity of 1.0 to itself).

also expect that the results would be strengthened if one considers tri- (or higher) connected components of the trust neighborhood. The Google API has been known to be filtering and restricting the number of the back-links it is reporting but it was the only tool available at the time of this research. Using the Yahoo Search API will likely improve the results we are getting.

V. CONCLUSIONS

In this paper we present a technique to identify spamming untrustworthy neighborhoods, developed by mimicking anti-propagandistic methods. In particular, we 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. Even though it may not be possible to identify spamming sites solely through our algorithm, our work is complementary to the recent developments that recognize web spam based on link analysis [45], [2].

One of the benefits of our method is that we do not need to explore the web graph explicitly in order to find these neighborhoods, which would be impossible for a client computer. Of course, it would be possible to support a user's trusted and untrusted sites through some personalization service provided by search engines. To be usable and efficient, this service would require the appropriate user interface. For example, a search engine's Toolbar could have a "Web Spam" button similar to the "Spam" or "Junk" buttons that many email applications fashion these days. When a user encounters an untrustworthy site coming high up in the results of some search query, she would select the item and click on a "Distrust" button. The browser would add this site in the user's untrustworthy site collection and would run the algorithm that propagates distrust backwards. Next time the user runs a similar search query, the untrusted sites would be blocked or demoted.

Recently, Google has introduced SearchWiki, a method of supporting personalized opinions about search results [41], which could be adjusted to support this operation. We view this development as justified by our findings and, even though we do not know whether Google's decision to employ this tool was partially influenced by our results, we do think it is a step in the right direction.

The algorithm we described is a first step in supporting the trust network of a user. 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. Organizations that the user joins and trusts may also add to this set. A combination of search engines capable of providing indexed content and structure [19], including identified neighborhoods, with personalized filtering those neighborhoods through the user's trust preferences, would provide a new level of reliability to the user's information gathering. Sharing ranking decisions with the end user will make it much harder for spammers to tune to a single metric – at least as hard as it is for propagandists to reach a large audience with a single trick.

ACKNOWLEDGMENT

The author's research was partially funded by a Brachman-Hoffman Fellowship.

The author would like to thank Joe DeStefano, Mirena Chausheva, Meredith Beaton-Lacoste, Scott Anderson and Scott Dynes for their valuable contributions. We would also like to thank David "Pablo" Cohn for his many useful suggestions that improved the presentation of the paper. The graphs shown in this paper were drawn using the yEd package [46].

REFERENCES

[1] 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.

[2] A. Benczúr, K. Csalogány, T. Sarlós, and M. Uher. Spam Rank – Fully automatic link spam detection. In *Proceedings of the AIRWeb Workshop*, May 2005.

[3] 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.

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

[5] 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.

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

[7] A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, and J. Wiener. Graph structure in the web. *Comput. Networks*, 33(1-6):309–320, 2000.

[8] J. Cho and S. Roy. Impact of search engines on page popularity. In *Proceedings of the thirteenth international conference on World Wide Web*, May 2004.

[9] CNETNews. Engine sells results, draws fire. http://news.cnet.com/2100-1023-215491.html, June 21 1996.

[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] 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.

[16] Z. Gyöngyi and H. Garcia-Molina. Web spam taxonomy. In *Proceedings of the AIRWeb Workshop*, May 2005.

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

[18] S. Hansell. Google keeps tweaking its search engine. New York Times, Jun. 3 2007.

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

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

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

[22] 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.

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

[24] 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.

[25] JUNG. The JUNG framework developer team – release 1.5. http://jung.sourceforge.net/.

[26] J. Kleinberg. The small-world phenomenon: an algorithm perspective. In *STOC '00: Proceedings of the thirty-second annual ACM symposium on Theory of computing*, pages 163–170. ACM Press, 2000.

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

[28] 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.

[29] R. Kumar, P. Raghavan, S. Rajagopalan, and A. Tomkins. The web and social networks. *IEEE Computer*, 35(11):32–36, 2002.

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

[31] 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.

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

[33] M. L. Maulding. Lycos: Design choices in an internet search service. *IEEE Expert*, January-February(12):8–11, 1997.

[34] P. Metaxas. On the evolution of search engine rankings. In *Proceedings of the 5th WEBIST Conference*, Lisbon, Portugal, March 2009.

[35] P. Metaxas. Using propagation of distrust to find untrustworthy web neighborhoods. In *Proceedings of the 4th International Conference on Internet and Web Applications and Services* (ICIW 2009), Venice, Italy, May 2009.

[36] A. Ntoulas, D. Fetterly, M. Manasse, and M. Najork. Detecting spam web pages through content analysis. In *World-Wide Web 2006*, May 2006.

[37] 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.

[38] P. Raghavan. Social networks: From the web to the enterprise. *IEEE Internet Computing*, 6(1):91–94, 2002.

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

[40] 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.

[41] The official Google blog. SearchWiki: Make search your own. http://googleblog.blogspot.com/2008/11/searchwiki-make-search-your-own.html, Nov. 20 2008.

[42] 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.

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

[44] D. Welch. Power of persuasion - propaganda. *History Today*, 49(8):24–26, 1999.

[45] B. Wu and B. Davison. Identifying link farm spam pages. In *Proceedings of the fourteenth international conference on World Wide Web*, May 2005.

[46] yWorks. yEd – java graph editor, v. 2.2.1. http://www.yworks.com/en/products_yed_about.htm.