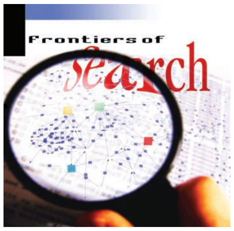
# Spam: It's Not Just for Inboxes Anymore



**E-mail spam is a nuisance that every user has come to expect. But Web spammers prey on unsuspecting users and undermine search engines by subverting search results to increase the visibility of their pages.**

*Zoltán Gyöngyi*
*Hector Garcia-Molina*
Stanford University

Despite the promises of software companies and service providers, e-mail spam is as familiar and annoying as the deluge of solicitations in snail mail and just as inevitable. It is often surprising and more than a little dismaying, however, when an innocent Web search for a local auto-body repair shop lands you on a page with great mortgage rates, inexpensive prescription drugs, legal assistance, or painfree cosmetic surgery. It is as if you have inadvertently clicked on all the worst links in your e-mail inbox, and who knows what is going on behind the scenes while you are on the unwanted page? Congratulations, you have just been Web spammed.

On the surface, such misdirection seems like blatant search engine abuse, but a closer look reveals a slightly fuzzier picture, as the "What Constitutes Web Spam?" sidebar describes. Despite these gray areas, Web spam *is* a growing problem.

The issue—and what makes this version more dangerous than e-mail spam—is that Web spam undermines the reputation of a trusted information source. The difference boils down to expectations. We can view all obvious sources of influence with varying degrees of skepticism. E-mail spammers don't undermine trust except in the very naive. And most of us view the advertisements on Web pages with a healthy dose of skepticism.

Web spamming, in contrast, undermines the trustworthiness people have come to expect from search engines. Google, Yahoo, and MSN have presented results that satisfy our information needs, and they have built a reputation on providing reliable, unbiased, trustworthy references. In short, we trust them and we have extended that trust to the results that they return. Contrast this to e-mail spam, which is largely a nuisance; we expect no value from it, so we delete it. But we tend to view search results as unbiased and trustworthy, so we click with confidence.

Web spammers are counting on this trust and that more people will turn to search engines for their information needs. Those in the search engine community believe that Web spam will become increasingly prevalent and sophisticated, and statistical data, although sparse, supports that view. Reports in 2002 indicated that about six to eight percent of the pages in a search engine index were spam,[1] while reports from 2003 to 2004 showed 15 to 18 percent.[2,3] Another study found that about nine percent of search results contain at least one spam link in the top-10 list, while 68 percent of all queries contain some spam in the top-200 list.[4]

Search engine companies are fighting back by inflicting penalties on obvious offenders, but users can also help by becoming more aware of the Web spammer's bag of tricks and looking more closely at search results. Thwarting the spammer's intent—to get hits on their pages—by ignoring the spamming link can help reverse this disturbing trend.

## MOTIVATION

A first step in fighting back is to understand the Web spammers' motives. Although some spamming is based on the desire to further political or religious interests, such as the "miserable failure" example in the sidebar, most spammers just want financial gain.

Some organizations fall into spam quite innocently, hiring search engine optimizers (SEOs) to

Published by the IEEE Computer Society 0018-9162/05/$20.00 © 2005 IEEE

make their Web assets more accessible. Many SEOs are completely legitimate in their Web page restructuring. They might improve visual design, add navigational links, or make the Web page more accessible to search engines, for example, by adding appropriate caption text to images that are otherwise hard to index. The trouble starts when less scrupulous SEOs engage in spamming to improve the target site's ranking beyond what its content value justifies. Many clients who hire SEOs are quite unaware of their spamming activity until some search engine company penalizes the optimized target site.

Web spammers also profit by participating in affiliate programs. Many online merchants, notably Amazon.com, have such programs. The author of Web page *p* sets up a link to a specific product purchase on page *q*. If a Web user then reaches page *q* by following the link on *p* and buys the product, page *p*'s author receives some part of the transacted amount (usually around 5 percent). Similarly, Web authors can place advertisements on their pages and receive some money when a visitor clicks on an advertisement.

All these programs represent an incentive for spammers, who construct (often automatically) Web pages that have at best minimal original content (some have none), but contain affiliate links. The spammer then attempts to boost the ranking of these pages to attract more visitors. Indeed, from our conversations with search engine experts and our evaluation of Web spam pages, affiliate programs appear to be the most common Web spamming trigger.

## SPAMMING TECHNIQUES

Search engines aim to provide high-quality results by correctly identifying all Web pages that are relevant for a specific query and then presenting the user with some of the more important ones. Most search engines determine a relevance score by measuring the textual similarity between the query and page. Importance is typically the page's popularity with respect to either all other pages on the Web (global, query-independent importance) or the other pages the engine has already identified as relevant. A common computational approach is to infer importance from the link structure—pages with more incoming links are more important, for example. Another approach is to maintain a log of results that users clicked on in previous search sessions and to use the clicks that each page received to gauge its importance.

Regardless of the specific method, a search engine produces a ranking score for each page that combines its relevance and importance to order the query results. Spamming techniques boost the ranking of specific pages by targeting algorithms that determine relevance and/or importance. Boosting works with hiding techniques, which attempt to conceal the telltale signs of boosting from users or search-engine experts. (Some readers might be concerned that by publishing these techniques we are aiding spammers, but we are merely reporting what is already in widespread use.)

### What Constitutes Web Spam?

Some Web spam is an obvious subversion. When we wrote this article, the top result of a Google search for "Kaiser pharmacy online" was a Web page that looked exactly like a Google result page, except that the header was "Gogle" and all the links pointed to a handful of sites selling cheap prescription drugs. Clearly, this crude result page is a deliberate attempt to undermine Google's reputation and undoubtedly deserves the "spam" label.

But consider another case: The Web site for World News Network ranks third on Google for the queries "world news" and "news network," in the company of well-known sources such as BBC, CNN, and Fox. To the trusting party making the query, that would seem logical except that WNN is not a news source; it's a company that owns several thousand Web sites, each apparently an online newspaper on specific topics in specific geographic regions (*Asia Maritime* and *Cairo Business*, for example). Further investigation of the sites reveals many articles that are out of context. Most of the articles in *Cairo Business*, for example, are only vaguely related to business news in Cairo, covering more general events such as the war in Iraq. The sites are tightly linked to other WNN sites, while no sites unaffiliated with WNN point to any of them. Although all the articles come from reputable sources like the Associated Press, CNN, and *The New York Times*, references to the actual sources are sparse. Thus, there's some gray area here. News aggregation can be useful, but does WNN provide something genuinely valuable, or does it exist mainly to collect revenue from the advertisements on its pages?

Adding to the fuzzy picture is the inherent subjectivity of the searcher and the whole freedom-of-expression issue. When we typed "miserable failure" in Google and Yahoo, the first or second hit was the official biography of President George W. Bush. (Michael Moore and Hillary Rodham Clinton were also high in the top 10.) A spamming technique did indeed affect the ranking, but this is not an obvious case of Web spamming because whether you cheer or jeer is a matter of individual preference. So even if the value is in the rather odd result, there is still value.

Thus, defining Web spam is not as straightforward as it might seem. On one hand, Web spamming is a blatant way to influence what people are exposed to. On the other hand, this is hardly a new problem. People face myriad attempts to sway them every day—friendly recommendations, telemarketing, television and newspapers, political propaganda, and evangelism. What one person deems obviously incorrect, misleading, frustrating, and annoying, others could find in alignment with their views or needs. Machine-generated gibberish—of no value to anyone—is easy to classify as spam. But what about that large gray area, where content has at least some value to someone? And who decides where the gray value area ends and the no-value realm begins?

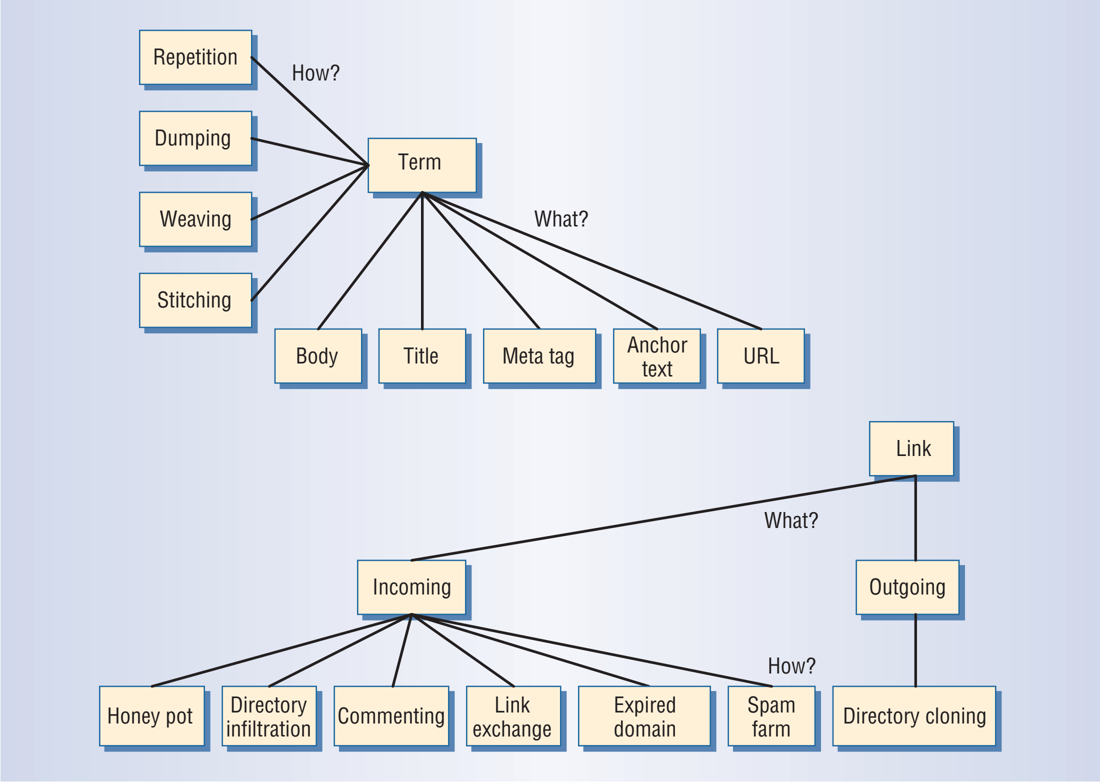


***Figure 1. Boosting techniques. In term spamming, the focus is on altering the text fields of a page (body, title, meta tags, anchor text, and URL), for example, by inserting terms that will promote the page for specific queries. In link spamming, the focus is on changing the link structure by adding links to (outgoing) or from (incoming) other Web pages.***

## Boosting

Spammers can boost a page ranking either by term spamming—editing a page's textual content—or link spamming—manipulating the link structure around that Web document.

**Term spamming.** In evaluating textual relevance, search engines consider various *fields* of Web pages, essentially the potential locations of the query terms. As Figure 1 shows, the most common text fields searched are the title, document body, the meta tags in the HTML header, the page's uniform resource locator (URL), and anchor text.

Figure 2 illustrates these text fields in the HTML code of a sample page about online gambling. As the figure shows, the search engine also considers the anchor text associated with URLs that point to the page as part of the page's text because the anchor text often accurately describes the page's content.

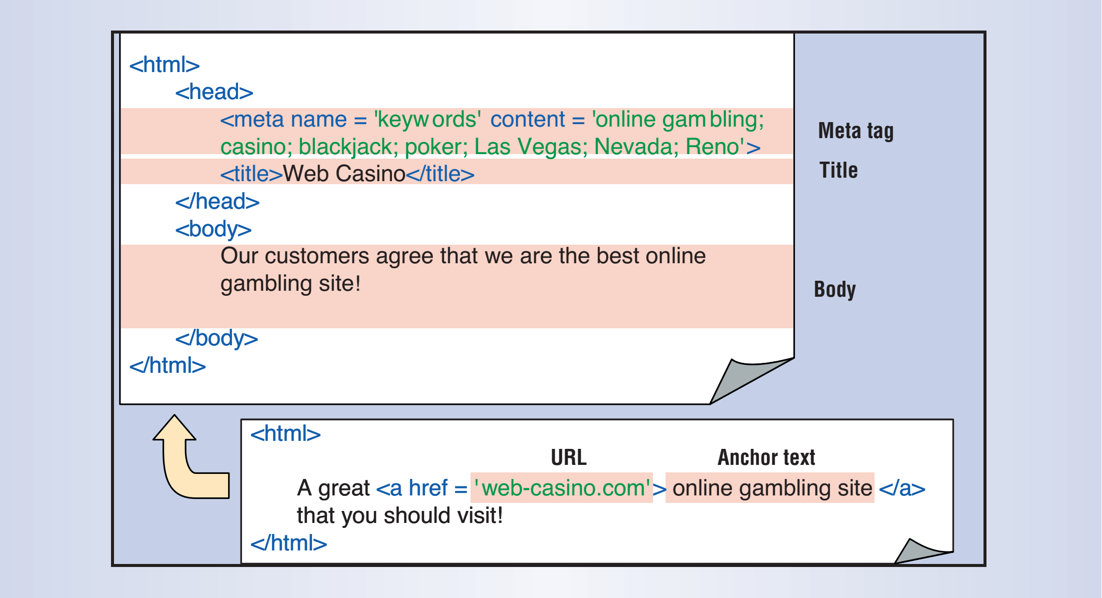
Search engine algorithms identify matches between query terms and field terms to determine the page's relevance score with respect to the query. Usually, a page is more relevant if query terms occur frequently and in proximity to each other.

The algorithms give different fields different weights according to their usefulness in determining relevance; for example, they typically weight anchor text more heavily than meta tags, which are very easy to spam. Anchor text spamming is also effective because a search engine indexes the anchor text for both the link source and target pages, so spamming affects the relevance of multiple pages. Moreover, the page's author typically has limited (and sometimes no) control of the anchor text, which means a spammer can negatively promote an adversarial Web page. The "miserable failure" example in the sidebar is an example of this.

Spammers can use several techniques to add terms to each text field. With the *repetition* technique, spammers repeat a few specific terms to increase the page's relevance with respect to those terms. By *dumping* a large number of unrelated terms, often even entire dictionaries, spammers can make a certain page relevant to many different queries. Dumping is effective if queries include relatively obscure terms because only a few pages are likely to be relevant. Even a spam page with low relevance or importance would be among the top results in that case.

*Weaving* spam terms into copied content is another way to mislead search engines, especially those that filter out plain repetition. In weaving, spammers duplicate bodies of news articles and online encyclopedia entries and insert spam terms randomly. Again, this ensures that the page is among the top results, particularly if the original text's topic was so rare that only a few relevant pages exist. In the following example, spam weaving conceals "airfare," "plane tickets," "cheap travel," "hotel rooms," and "vacation" in a quote from Benjamin Franklin:

> Remember not only airfare to say the right plane tickets thing in the right place, but far cheap travel more difficult still, to leave hotel rooms unsaid the wrong thing at vacation the tempting moment.



```
<html>
    <head>
        <meta name = 'keywords' content = 'online gambling;
        casino; blackjack; poker; Las Vegas; Nevada; Reno'>
        <title>Web Casino</title>
    </head>
    <body>
        Our customers agree that we are the best online
        gambling site!
    </body>
</html>
```

```
<html>
    A great <a href = 'web-casino.com'> online gambling site </a>
    that you should visit!
</html>
```

**Figure 2. Five text fields that search engines evaluate for relevance to a set of query terms. The meta tag, title, and body text fields are part of the page itself (in this case, an online gambling page), while the URL and anchor text are external. Search engines store and use these external fields along with the actual page content because they tend to offer concise descriptions of the page content.**

*Phrase stitching* is another way to create content quickly. The idea is to combine sentences or phrases, possibly from different sources, such as really simple syndication (RSS) feeds. The spam page might then show up for queries on any of the topics in the original sentences. A spammer using this article as a source might come up with the following meaningless collage:

> Search engines aim to provide high-quality results by correctly identifying. Weaving conceals "airfare," "hotel rooms" in a quote from Benjamin Franklin. The research community has started investigating some common features on blogs.

**Link spamming.** In this variation of boosting, spammers create link structures they hope will increase the importance of one or more of their pages. Search engines typically use PageRank[5] or the Hubs and Authorities (H+A) algorithm[6] to assign importance scores based on link structure. Although both algorithms are somewhat spam resistant, spammers can subvert either one by creating groups of strongly interconnected pages or accumulating links from many potentially well-known, reputable Web pages (H+A is also sensitive to adding links *pointing to* well-known pages).

As Figure 1 shows, these are the two main categories in link spamming. In the *outgoing link* approach, the spammer adds outgoing links to a page in the hope that H+A will award that page a higher importance score. The most widespread method for doing this is *directory cloning*. The Web has many directory sites: Some are larger and better known, such as the DMOZ Open Directory (dmoz.org) or the Yahoo directory (dir.yahoo.com); others are smaller and less famous, such as the Librarian's Index to the Internet (lii.org). These directories organize Web content around topics and subtopics and list relevant sites for each. By replicating some or all of the directory pages, spammers can quickly create new groups of pages with a massive number of outgoing links.

In the *incoming link* approach, the idea is to accumulate many incoming links to a single target page or set of pages. Here, spammers have a wider choice of strategies. One is to create a *honey pot*, a set of pages that provide some useful resource, such as copies of some Unix documentation pages, but also have hidden links to one or more spam pages. The honey pot then attracts users to point to it, indirectly boosting the spam page's ranking. Directory clones could also act as honey pots.

Another strategy is to *infiltrate a Web directory*. Several Web directories let Web masters post links to their sites under some topic in the directory. In some cases, those who edit such directories do not strictly control and verify link additions. In others, the spammer is skilled enough to circumvent censorship and include spam links into the directories. Because directories tend to have high PageRank and H+A scores, spammers can use directory infiltration to tamper with the rankings that both algorithms produce.

Some spammers *post links* on blogs, unmoderated message boards, guest books, or wikis, including URLs to their spam pages as part of seemingly innocent comments or messages. Without an editor or a moderator to oversee all submissions, a blog, message board, or guest book is vulnerable to spam. Even with an editor or a moderator, detecting spam comments could be difficult, particularly if the spammer uses a *hiding technique*. For example, in

> Nice story. Read about my <a href="http://bestcasinoonlinever.com">Las Vegas casino</a> trip.

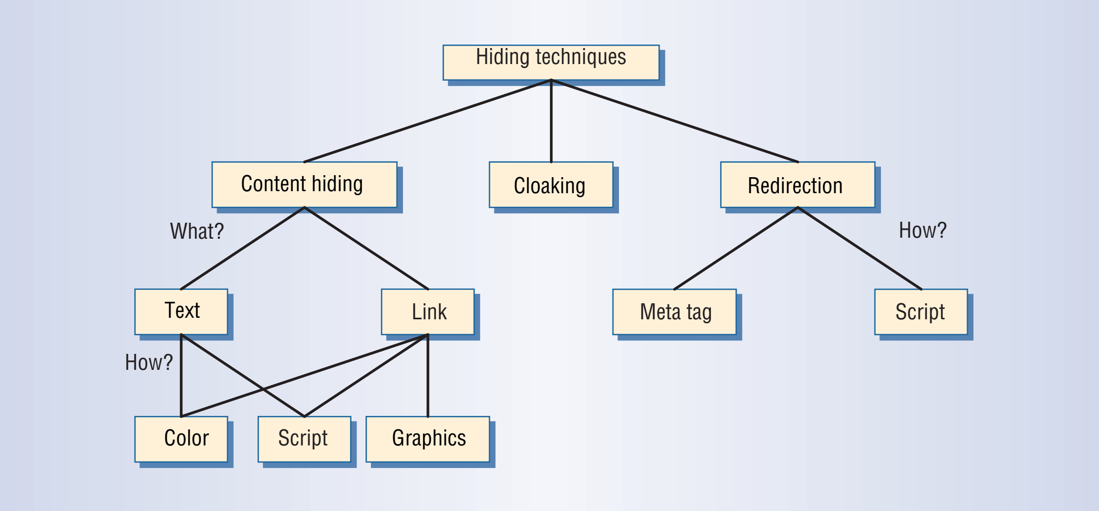


***Figure 3. Hiding techniques. Content hiding conceals text or links by using color schemes, scripts, or graphics. Cloaking is an attempt to deceive Web crawlers, while redirection sends the browser to another URL as soon as it loads the page.***

the spammer has used both link and anchor text spamming. The reader might see only the seemingly innocent "Nice story. Read about my Las Vegas casino trip" and not learn about the real intent until after clicking on the link.

Spammers know they aren't alone in their pursuits, and often they will set up a *link exchange structure* so that their sites point to each other. The hope is that this reciprocity will boost the importance of all the participants' pages.

When a domain name expires, the URLs on other Web sites might continue to point to pages within the expired domain. Spammers can *buy expired domains* and populate them with spam that exploits the relevance or importance of the old links.

Finally, creating a large *spam farm*—a group of sites with a link structure that boosts the ranking of some target pages—has become affordable. This approach, prohibitively expensive only a few years ago, has become more common since the costs of domain registration and Web hosting have dramatically declined.

## Hiding techniques

Some boosting techniques leave considerable evidence on Web pages, such as an abundance of links or unusually long anchor text phrases. Spammers often try to conceal these telltale signs by using hiding techniques that make the page more appealing to visitors, whether those are Web browsers or search engine experts, who must dig beyond a simple visual inspection of the page to produce some proof of spamming.

Figure 3 details three main hiding techniques—content hiding, cloaking, and redirection—that spammers use to try to fool both Web browsers and search engines.

**Content hiding.** In this technique, parts of a page become invisible when a browser displays that page. The oldest technique is to match text and page background color using cascading stylesheets or appropriate HTML tags:

```
<body background="white">
   <font color="white">hidden
   text</font>
…
</body>
```

Similar techniques can hide anchor text as well. Spammers commonly create tiny (say 1 × 1 pixel) anchor images that either are transparent or have the background color. They also use scripts to hide some of a page's content by setting an HTML element's (such as a paragraph) visible style attribute to false, for example.

**Cloaking.** If spammers can clearly identify the Web crawlers that search engines use, they can adopt cloaking. In this strategy, for the same URL, spam servers return one HTML document to a regular Web browser and a different document to a Web crawler. Spammers can thus present the intended content to Web users with no spam on the page while simultaneously sending a spammed document to the search engine for indexing.

Some spammers identify Web crawlers by sifting through a list of Internet Protocol (IP) addresses or domain names that they know search engines commonly use, such as for Google, googlebot.com. Others have their servers look at the user-agent field in the HTTP request message. In the following simple HTTP request message, the user-agent name is one that a version of the AltaVista crawler uses:

```
GET/db_pages/members.html
HTTP/1.0
Host: infolab.stanford.edu
User-Agent: AVSearch-
3.0(AltaVista/AVC)
```

User-agent names are not strictly standard; the requesting application can decide what to include in the corresponding message field. Nevertheless, search engine crawlers identify themselves by a

**Table 1. Possible antispam solutions.**

| Proposed solution | Spam targeted | Function | Automatic? | Useful for |
|---|---|---|---|---|
| Statistical language model | Term spamming, in particular blog infiltration | Identifies unnatural word distribution | Yes | Spam detection |
| Analysis of link-count distribution | Link spam farms | Looks at in-degree and out-degree distribution outliers | Yes | Spam detection |
| Analysis of PageRank distribution | Link spam farms | Looks for unnatural PageRank score distributions of in-neighbor pages | Yes | Spam detection |
| Collusion detection | Link spam farms | Identifies groups of strongly interconnected pages | Yes | Rendering specific spamming technique ineffective |
| TrustRank | All types | Separates reputable pages from spam on the basis of connectedness to a set of known reputable pages (seed) | Semi (requires manually compiled seed set) | Spam demotion |

name distinct from the ones that traditional Web browsers use. Webmasters are thus free to block access to some content, control network traffic parameters, or perform well-intended, legitimate optimizations.

**Redirection.** In this strategy, the spam server automatically redirects the Web browser to another URL as soon as the page loads. This way the search engine still indexes the page, but the user never sees it. Pages with redirection are in essence intermediates (proxies or doorways) for the ultimate targets, which spammers try to serve to users through search engines.

A simple redirection approach is to take advantage of the refresh meta tag in the HTML document's header. By setting the refresh time to zero and the refresh URL to the target page, spammers can redirect the page as soon as the browser loads it:

```
<meta http-equiv="refresh"
content="0;url=target.html">
```

Implementing this approach is relatively easy, but search engines can easily identify such redirection attempts by parsing the meta tags. More sophisticated spammers achieve redirection as part of a script on the page, since crawlers are unlikely to execute the scripts:

```
<script language="javascript"><!--
  location.replace("target.html")
--></script>
```

## THE DETECTION ARSENAL

Perhaps the most straightforward way to detect Web spam is manually. Search engines used to rely on a handful of human editors who swept through the index, identified spam pages, and penalized them according to some well-established policy. For a time, this approach alone was good enough, but it rapidly became too expensive and limited to accommodate Web growth. Search engines have since moved to algorithmic spam detection that requires little or no human intervention.

Fearing new waves of spamming, search engine companies have been reluctant to discuss their ranking algorithms or spam detection techniques; hence, much secrecy surrounds the practice of counteracting spam. However, because Web spam is becoming more of a threat, the research community has recently started investigating it, yielding the handful of possible solutions in Table 1.

Algorithms can use statistical language models to identify instances of term spamming because the word distribution of a spam page frequently differs from that of everyday written natural language. Algorithms can also identify spam by detecting discrepancies between the word distribution of spam content and of the context (surrounding text or pages). Researchers have used the latter approach successfully to detect spam in blog comments.[7]

As the table shows, several solutions target link spam. Large spam farms are often machine-generated and consequently have regular structures that are not hard to identify. Dennis Fetterly and colleagues[1] analyzed the incoming and outgoing link count (in-degree and out-degree) distributions of Web pages. These distributions typically follow a power-law pattern—only a few Web pages have a large in- or out-degree, while most documents have only a few incoming or outgoing links. Occasionally, however, search engines encounter substantially more pages with a specific in- or out-degree than what the distribution formula predicted. The vast majority of such outliers are pages that belong to large spam farms.

When a reputable Web page $p$ has many incoming links, the power-law formula also applies to the PageRank scores of the pages pointing to $p$. The target page $q$ of a large spam farm tends to have many incoming links as well. However, the machine-generated spam pages pointing to $q$ share the same or very similar PageRank score. Therefore, checking the PageRank distribution of linked pages helps detect spam, as András Benczúr and colleagues have shown.[3]

Another group of link-spam detection techniques focuses on heavily interlinked pages. Collusion is an efficient way for spammers to improve PageRank or H+A scores. Hui Zhang and colleagues[8] and Baoning Wu and Brian Davison[4] have proposed efficient algorithms for collusion detection. However, collusion alone is not a guarantee of spamming because reputable pages are also often strongly interconnected. Therefore, collusion detection is best used for penalizing all suspicious pages during ranking, as opposed to reliably pinpointing spam.

Other research looks at how to address the spamming problem globally by identifying common features of spam pages, without targeting a particular spamming technique. One such antispam solution, TrustRank,[2] is based on the *approximate isolation* of nonspam pages. Because reputable Web pages seldom point to spam, the idea is to start with some reputable pages (taken from a list that an expert has compiled manually) and apply an algorithm that circumspectly propagates the quality judgment to (directly or indirectly) connected pages. In this way, it is possible to separate reputable pages from spam, promoting the former group in ranking while demoting the latter.

Web spamming has far-reaching technical, economic, and social implications. Because it is not always clear what constitutes spam and because the sophistication of spammers is constantly increasing, no single solution on the horizon will eradicate spam. Fighting spam is an ongoing battle: the more advanced the spam detection techniques, the more sophisticated the spammers.

Search engine companies will continue to fight back by keeping spammers in the dark about their antispam methods. But in the long run, the best solution to the ongoing battle is to make spamming ineffective—not only in its attempt to subvert search engine algorithms but also—and more important—in its attempt to coerce users. If people are more conscious about spamming and avoid being lured into its traps, the economic or social incentive for spamming will decrease. Hopefully, the techniques we have described will help in both these missions. ■

## References

1. D. Fetterly, M. Manasse, and M. Najork, "Spam, Damn Spam, and Statistics," *Proc. Int'l Workshop on the Web and Databases* (WebDB), ACM Press, 2004, pp. 1-6.
2. Z. Gyöngyi, H. Garcia-Molina, and J. Pedersen, "Combating Web Spam with TrustRank," *Proc. Int'l Conf. Very Large Databases* (VLDB), Morgan Kaufmann, 2004, pp. 576-584.
3. A. Benczúr et al., "SpamRank—Fully Automatic Link Spam Detection," *Proc. Int'l Workshop Adversarial Information Retrieval on the Web* (AIRWeb), 2005; http://airweb.cse.lehigh.edu/2005/#proceedings.
4. B. Wu and B. Davison, "Identifying Link Farm Spam Pages," *Proc. Int'l Conf. World Wide Web*, ACM Press, 2005, pp. 820-829.
5. L. Page et al., *The PageRank Citation Ranking: Bringing Order to the Web*, tech. report, Computer Science Dept., Stanford Univ., 1998.
6. J. Kleinberg, "Authoritative Sources in a Hyperlinked Environment," *J. ACM*, vol. 46, no. 5, 1999, pp. 604-632.
7. G. Mishne, D. Carmel, and R. Lempel, "Blocking Blog Spam with Language Model Disagreement," *Proc. Int'l Workshop Adversarial Information Retrieval on the Web* (AIRWeb), 2005; http://airweb.cse.lehigh.edu/2005/#proceedings.
8. H. Zhang et al., "Making Eigenvector-Based Reputation Systems Robust to Collusion," *Proc. Int'l Workshop Algorithms and Models for the Web-Graph* (WAW), Springer-Verlag, 2004, pp. 92-104.

***Zoltán Gyöngyi*** *is a PhD student in the Computer Science Department at Stanford University. His current research interest is Web search. Gyöngyi completed the DiplEng program in computer science and engineering at the Technical University of Cluj-Napoca, Romania. Contact him at zoltan@cs.stanford.edu.*

***Hector Garcia-Molina*** *is a professor in the Computer Science and Electrical Engineering Departments at Stanford University. His research interests include peer-to-peer systems, entity resolution, and Web search. Garcia-Molina received a PhD in computer science from Stanford University. He is a member of the ACM and the American Academy of Arts and Sciences. Contact him at hector@cs.stanford.edu.*