Countering Web Spam Using Link-Based Analysis

James Caverlee Mudhakar Srivatsa Ling Liu

College of Computing
Georgia Institute of Technology
Atlanta, GA 30332 USA
{caverlee, mudhakar, lingliu}@cc.gatech.edu

Web spam refers to efforts by malicious adversaries to manipulate how users view and interact with the World Wide Web, often to drive traffic to particular spammed Web pages, regardless of the merits of those pages. As the Web has grown and increasingly become the primary platform for information sharing and electronic commerce, there has been a rise in targeted Web spam that is designed to degrade the quality of legitimate Web sites (and the services they offer) and to manipulate the user experience for the advantage of the Web spammer. In particular, we identify three major categories of Web spam:

- **Page Spoofing:** To support identify theft, Web spammers often construct illegitimate copies of legitimate Web sites (like `www.ebay.com`). Users are then directed to these spoofed sites through email-based phishing attacks or spammer-controlled fake Web directories.
- **Browser-Based Attacks:** Browser-based spam includes techniques that directly attack the Web browser technology for the gain of the Web spammer; for example, the browser may display a legitimate hyperlink that when clicked is replaced by an alternative spammed hyperlink.
- **Search Engine Manipulation:** Since search engines play such a central role in bringing the top-matched Web pages to the vast majority of Web users, a considerable amount of malicious Web spamming is focused on manipulating the ranking algorithms that drive search engines.

Ultimately, all three types of Web spam (i) degrade the quality of information on the Web, resulting in the pollution of search engines' indexes and the worsening of the end user Web experience; and (ii) place the user at risk for further malicious exploitation by the Web spammer. Web spam is a serious problem, and recent studies suggest that it accounts for a significant portion of all Web content, including 8% of pages [1] and 18% of sites [2]. We anticipate that any successful effort to resist all forms of Web spam will rely on a suite of approaches. In the rest of this abstract, we focus our attention on the problem of search engine manipulation.

With respect to search engines, the biggest Web spam problem is link-based spam because it attacks the link-based ranking algorithms behind search engines. Link-based ranking is a very important topic – prominent examples include the query-dependent HITS algorithm [3] and the query-independent PageRank algorithm for assigning a global "authority" score to each page on the Web [4]. Since link-based ranking approaches are at the core of modern search engines and the subject of considerable attention by Web spammers, we focus our efforts on developing robust countermeasures to link-based spam. We identify three prominent types of link-based Web vulnerabilities:

- The hijacking of links from high-quality pages so that reputable pages appear to endorse spam pages;
- Collusion strategies among multiple pages, whereby spammers create complex link exchange ar-

rangements to outwit link-based ranking algorithms; and

- Link boosting arrangements like link farms that rely on brute-force link manipulation through the addition of thousands or millions of dummy pages all linking to a target spam page.

Each of these link-based attacks subvert the credibility of traditional link-based ranking approaches and undermine the quality of information offered through search engines. To defend against these three important types of link-based vulnerabilities, we have developed a suite of targeted countermeasures. Of course any approach to deterring Web spam is faced with the the classic arms race cycle endemic to security-related research, that is: (i) a solution is proposed; (ii) the spammers adapt their techniques to subvert the solution; (iii) the solution is revised, the spammers adapt, and the cycle continues. Our targeted countermeasures are designed to significantly raise the costs of link-based manipulation, so that Web spammers wield only a limited ability to impact link-based algorithms and to continue the arms race cycle.

One such countermeasure we have developed relies on a notion of *hijack-resistant influence flow* to selectively throttle the influence of Web spammers. The countermeasure limits the impact of hijacked pages from sources outside of the complete control of the Web spammer, so that it is more difficult for spammers to capture endorsements from reputable pages. We incorporate this countermeasure into a PageRank-style iterative algorithm that relies on a source view of the Web. This "SourceRank" approach assigns a score to each page based on the overall quality of the source that the page belongs to through a random walk over Web sources. Since SourceRank considers the relative merits of logical collections of Web pages, it can provide more robust Web rankings, making it harder for adversaries to take advantage of the ranking system. Analytically, we provide a formal discussion on the effectiveness of the countermeasure-strengthened SourceRank approach against link-based Web spam. Experimentally, we show how the proposed countermeasure provides strong resistance to manipulation and significantly raises the cost of rank manipulation to a Web spammer. We additionally provide a detailed formal analysis of the spam resilience properties of the proposed countermeasure, and validate the findings of this analysis over real-world Web data of over 170 million pages.

References

[1] D. Fetterly, M. Manasse, and M. Najork. Spam, damn spam, and statistics. In *WebDB*, 2004.

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

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

[4] L. Page et al. The PageRank citation ranking: Bringing order to the Web. Technical report, Stanford, 1998.