Spam Double-Funnel: Connecting Web Spammers with Advertisers

Yi-Min Wang, Ming Ma
Microsoft Research
Redmond, WA 98052, USA
425-882-8080
{ymwang, mingma}@microsoft.com

Yuan Niu, Hao Chen
University of California, Davis
Davis, CA 95616-8562, USA
530-754-5375
{niu, hchen}@cs.ucdavis.edu

ABSTRACT

Spammers use questionable search engine optimization (SEO) techniques to promote their spam links into top search results. In this paper, we focus on one prevalent type of spam – redirection spam – where one can identify spam pages by the third-party domains that these pages redirect traffic to. We propose a five-layer, double-funnel model for describing end-to-end redirection spam, present a methodology for analyzing the layers, and identify prominent domains on each layer using two sets of commercial keywords – one targeting spammers and the other targeting advertisers. The methodology and findings are useful for search engines to strengthen their ranking algorithms against spam, for legitimate website owners to locate and remove spam doorway pages, and for legitimate advertisers to identify unscrupulous syndicators who serve ads on spam pages.

Categories and Subject Descriptors

H.3.5 [**Information Storage and Retrieval**]: Online Information Services - *Commercial services, Web-based services*

General Terms: Measurement, Security, Experimentation

Keywords: Search Spam, Web Spam, Redirection and Cloaking, Advertisement Syndication

1. INTRODUCTION

Search spammers (or ***web spammers***) refer to those who use questionable search engine optimization (SEO) techniques to promote their low-quality links into top search rankings. Common SEO techniques include stuffing keywords, creating link farms (e.g., large number of mutually linked, made-for-ads websites), posting links to spam pages as comments at public forums (referred to as *comment spamming*), and using ***crawler-browser cloaking*** techniques [8] to serve different pages to crawlers and end users. To evade spam investigation, some spammers in recent years have started using ***click-through cloaking*** techniques [15,22] to display bogus content to spam investigators who visit their pages directly without clicking through any search results.

We use ***redirection spam*** to refer to the web pages that redirect browsers to visit known spammer-controlled third-party domains. Many redirection spam pages use ***syndication*** where they participate in pay-per-click programs and display ***ads-portal pages***.

WWW 2007, May 8–12, 2007, Banff, Alberta, Canada.
ACM 978-1-59593-654-7/07/0005.

We motivate our work using a real example. Around mid-October 2006, the following three *doorway* URLs appeared among the top-10 Live Search results for "cheap ticket":

- http://-cheapticket.blogspot.com/
- http://sitegtr.com/all/cheap-ticket.html
- http://cheap-ticketv.blogspot.com/

All these pages appeared to be spam: they used cloaking, their URLs were posted as comments at numerous open forums[1], and they redirected traffic to known-spammer redirection domains vip-online-search.info, searchadv.com, and webresourses.info. Surprisingly, ads for orbitz.com, a reputable company, appeared on all these three spam pages. A search using similar keywords[2] at Google and Yahoo! revealed another two spam pages, hosted on hometown.aol.com.au and megapage.de, that also displayed orbitz.com ads. If we believe that a reputable company is unlikely to buy service directly from spammers, a natural question to ask is: *who are the middlemen who indirectly sell spammers' service to sites like orbitz.com?*

We discovered the answer by ***"following the money"***: when we clicked the orbitz.com ads on each of the five pages and monitored the resulting HTTP traffic using the Fiddler tool [27], we saw that the ads click-through traffic got *funneled* into either 64.111.210.206 or the block of IP addresses between 66.230.128.0 and 66.230.191.255 [30]. Moreover, the chain of redirections stopped at http://r.looksmart.com, which then redirected to orbitz.com using HTTP 302.

In this paper, we analyze end-to-end redirection spam activities comprehensively with an emphasis on syndication-based spam. We propose a ***five-layer double-funnel model*** in which displayed ads flow in one direction and click-through traffic flows in the other direction. By constructing two different benchmarks of commercial search terms and using the ***Strider Search Ranger*** system [21] to analyze tens of thousands of spam links that appeared in top results across three major search engines, we identified the major domains in each of the five layers and their interesting characteristics.

The paper is organized as follows. Section 2 gives an overview of the Search Ranger system and introduces the double-funnel model. In Section 3 we construct a spammer-targeted search benchmark. Section 4 analyzes spam density and double-funnel for this benchmark. In Section 5 we construct an advertiser-

[1] For ease of presentation, throughout the paper, we use the term "forums" to include all blogs, bulletin boards, message boards, guest books, web journals, diaries, galleries, archives, etc. that can be abused by web spammers to promote spam URLs.

[2] We use the terms "keyword", "query", and "search term" interchangeably in this paper to refer to the entire query phrase that a user enters into a search box to perform a query.

targeted benchmark and compare the analysis results using this benchmark with those in Section 4. Section 6 discusses non-redirection spam that also connects to the double-funnel model. Section 7 surveys related work, and Section 8 concludes the paper. Since all the analyses in this paper are based on the data gathered in September and October of 2006, some spam URLs may no longer be active.

2. REDIRECTION SPAM

2.1 Definitions: Search Spam and Redirection

SEO techniques span a wide spectrum. Since the precise boundary between legitimate SEO techniques and search spam is often subjective and fuzzy, we focus on one type of spam – *redirection spam* – which is widely used by large-scale spammers to associate many doorway pages with a single redirection domain. These doorway pages often exhibit similar patterns in their appearance, their cloaking and code obfuscation techniques for avoiding detection, and the way by which their URLs appear in the comment fields of public forums. These repeated patterns allow human investigators to judge spam pages more easily and confidently. We will describe the exact steps in detecting spam in the next subsection. In Sections 4 and 5, we will show that redirection spam accounts for significant spam densities in both our benchmarks, which indicate that our spam detection mechanism is effective in practice.

After a user instructs the browser to visit a URL (the primary URL), the browser may visit other URLs (secondary URLs) automatically. The secondary URLs may contribute to inline contents (e.g., Google AdSense ads) on the primary page, or may replace the primary page entirely (i.e., they replace the URL in the address bar). We consider both these types of secondary URLs *redirection*. See [31] for screenshots of sample redirection spam.

2.2 Strider Search Ranger System

The Strider Search Ranger system [21] is an automated spam detection system with the following three key features:

1. Web Patrol with Search Monkeys [19] **-** Since search engine crawlers typically do not execute scripts, spammers exploit this fact using crawler-browser cloaking techniques, which serve one page to crawlers for indexing but display a different page to browser users [8,23]. To defend against cloaking, Search Monkeys visit each web page with a full-fledged popular browser, which executes all client-side scripts. To combat the newer click-through cloaking technique, which serves spam content only to users who click through search results, our monkey programs mimic the click-through by first retrieving a search-result page to set the browser's document.referrer variable, then inserting a link to the spam page in the search-result page, and finally clicking through the inserted link.

2. Follow the Money through Redirection Tracking – Common approaches to detecting "spammy" content and link structures merely catch ***"what"*** spammers are doing today. By contrast, if we follow the money by tracking traffic redirection, we would be closer to identifying ***"who"*** are behind spam activities, even if their spam techniques evolve. Search Ranger uses the *Strider URL Tracer* [20] to intercept browser redirection traffic at the network layer to record all redirection URLs. As Sections 4 and 5 will demonstrate, we apply redirection analysis to tracking both the ads-fetching traffic and the ads click-through traffic.

3. Similarity-based Grouping for Identifying Large-scale Spam – Rather than analyzing all crawler-indexed pages, Search Ranger focuses on *monitoring search results of popular queries targeted by spammers* to obtain a list of URLs with high spam densities. It then analyzes the similarity between the redirections from these pages to identify related pages, which are potentially operated by large-scale spammers. In its simplest form, this similarity analysis identifies doorway pages that share the same redirection domain. After we verify that the domain is responsible for serving the spam content, we then use the domain as a seed to perform "backward propagation of distrust" [13] to detect other related spam pages.

In summary, Search Ranger identifies spam URLs using the process summarized below.

Search Ranger Spam Detection Process

Step 1: Given a set of search terms and a target search engine, Search Monkeys retrieve the top-N search results for each query, remove duplicates, and scan each unique URL to produce an XML file that records all URL redirections.

Step 2: At the end of a batched scan, Search Ranger applies redirection analysis to all the XML files to classify URLs that redirected to known-spammer redirection domains as spam.

Step 3: Search Ranger groups unclassified URLs by each of the third-party domains that received redirection traffic.

Step 4: Search Ranger submits sample URLs from each group to a spam verifier, which gathers evidence of spam activities associated with these URLs. Specifically, the spam verifier checks if each URL uses crawler-browser cloaking to fool search engines or uses click-through cloaking to evade manual spam investigation. It also checks if the URL has been widely comment-spammed at public forums.

Step 5: Search Ranger submits groups of unclassified URLs, ranked by their group sizes and tagged by spam evidence, to human judges. Once the judges determine a group to be spam, Search Ranger adds the redirection domains responsible for serving the spam content to the set of known spam domains, which will be used in Step-2 classification in future scans.

2.3 Spam Double-Funnel

A typical advertising syndication business consists of three layers: the *publishers* who attract traffic by providing quality content on their websites to achieve high search rankings, the *advertisers* who pay for displaying their ads on those websites, and the *syndicators* who provide the advertising infrastructure to connect the publishers with the advertisers. The Google AdSense program [29] is an example syndicator. Although some spammers have abused the AdSense program [28], the abuse is most likely the exception rather than the norm.

In a questionable advertising business, spammers assume the role of publishers, who set up websites of low-quality content and use black-hat SEO techniques to attract traffic. To better survive spam detection and blacklisting by search engines, many spammers have split their operations into two layers. At the first layer are the ***doorway pages***, whose URLs the spammers promote into top search results. When users click those links, their browsers are instructed to fetch spam content from ***redirection domains***, which occupy the second layer.

To attract prudent legitimate advertisers who do not want to be too closely connected to the spammers, many syndicators have

also split their operations into two or more layers, which are connected by multiple redirections, to obfuscate the connection between the advertisers and the spammers. Since these syndicators are typically smaller companies, they often join forces through traffic aggregation to attract sufficient traffic providers and advertisers.

We model this end-to-end search spamming business with the five-layer double-funnel illustrated in Figure 1: tens of thousands of ***advertisers*** (Layer #5) pay a handful of ***syndicators*** (Layer #4) to display their ads. The syndicators buy traffic from a small number of ***aggregators*** (Layer #3), who in turn buy traffic from web spammers to insulate syndicators and advertisers from spam pages. The spammers set up hundreds to thousands of redirection domains (Layer #2), create millions of doorway pages (Layer #1) that fetch ads from these redirection domains, and widely spam the URLs of these doorways at public forums. If any such URLs are promoted into top search results and are clicked by users, all click-through traffic is funneled back through the aggregators, who then de-multiplex the traffic to the right syndicators. Sometimes there is a chain of redirections between the aggregators and the syndicators due to multiple layers of traffic affiliate programs, but almost always one domain at the end of each chain is responsible for redirecting to the target advertiser's website.

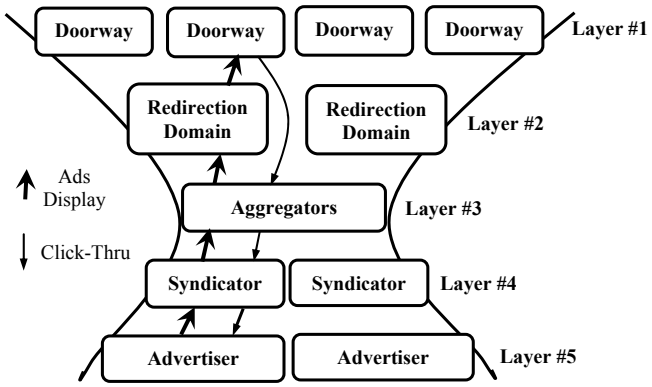

Figure 1: Spam Double-Funnel

In the case of AdSense-based spammers, the single domain googlesyndication.com plays the role of the middle three layers, responsible for serving ads, receiving click-through traffic, and redirecting to advertisers. Specifically, browsers fetch AdSense ads from the redirection domain googlesyndication.com and display them on the doorway pages; ads click-through traffic goes into the aggregator domain googlesyndication.com before reaching advertisers' websites.

3. SPAMMER-TARGETED KEYWORDS

To study the common characteristics of redirection spam, our first step was to discover the keywords and categories heavily targeted by redirection spammers. In this section, we describe our methodology for deriving 10 spammer-targeted categories and a benchmark of 1,000 keywords, which serve as the basis for the analyses presented in Section 4.

Redirection spammers often use their targeted keywords as the anchor text of their spam links at public forums, exploiting a typical algorithm by which common search engines index and rank URLs. For example, the anchor text for the spam URL http://coach-handbag-top.blogspot.com/ is typically "coach handbag". Therefore, we collect spammer-targeted keywords by extracting all the anchor text from a large number of spammed forums and ranking the keywords by their frequencies.

Between June and August of 2006, we manually investigated spam reports from multiple sources including search user feedback, heavily spammed forum types, online spam discussion forums, etc. We compiled a list of 323 keywords that returned spam URLs among the top 50 results at one of the three major search engines. We then queried these keywords at all three search engines, extracted the top-50 results, scanned them with an earlier version of Search Ranger, and identified 4,803 unique redirection-spam URLs.

Next, we issued a "*link:*" query on each of the 4,803 URLs and retrieved 35,878 unique pages that contained at least one of these spam URLs. From these pages, we collected a total of 1,132,099 unique keywords, with a total of 6,026,699 occurrences, and ranked the keywords by their occurrence counts. The top-5 keywords are all drugs-related: "***phentermine***" (8,117), "***viagra***" (6,438), "***cialis***" (6,053), "***tramadol***" (5,788), and "***xanax***" (5,663). Among the top one hundred, 74 are drugs-related, 16 are ringtone-related, and 10 are gambling-related.

Among the above 1,132,099 keywords, we could select a top list, say top 1000, for our subsequent analyses. However, we observed that keywords related to drugs and ringtones dominate the top-1000 list. Since it would be useful to study spammers who target different categories, we decided to construct our benchmark by manually selecting ten of the most prominent categories from the list. They are:

1. **Drugs**: phentermine, viagra, cialis, tramadol, xanax, etc.
2. **Adult**: porn, adult dating, sex, etc.
3. **Gambling**: casino, poker, roulette, texas holdem, etc.
4. **Ringtone**: verizon ringtones, free polyphonic ringtones, etc.
5. **Money**: car insurance, debt consolidation, mortgage, etc.
6. **Accessories**: rolex replica, authentic gucci handbag, etc.
7. **Travel**: southwest airlines, cheap airfare, hotels las vegas, etc.
8. **Cars**: bmw, dodge viper, audi monmouth new jersey, etc.
9. **Music**: free music downloads, music lyrics, 50 cent mp3, etc.
10. **Furniture**: bedroom furniture, ashley furniture, etc.

We then selected the top-100 keywords from each category to form our first benchmark of 1,000 spammer-targeted search terms.

4. REDIRECTION-SPAM ANALYSIS

In late September 2006, we submitted the 1,000 keywords to the Search Ranger system, which retrieved the top-50 results from all three major search engines. In total, we collected 101,585 unique URLs from 1,000x50x3=150,000 search results. With a set of approximately 500 known-spammer redirection domains and AdSense IDs at that time, the system identified ***12,635*** unique spam URLs, which accounted for ***11.6%*** of all the top-50 appearances. (The actual redirection-spam density should be higher because some of the doorway pages had been deactivated, which were no longer causing URL redirections when we scanned them.) We first give a brief analysis of per-category spam densities in Section 4.1 and then focus on the double-funnel analysis for the remainder of this section.

4.1 Spam Density Analysis

Figure 2 compares the per-category spam densities across the 10 spammer-targeted categories. The numbers range from ***2.7%***

for Money to ***30.8%*** for Drugs. Two categories, *Drugs* and *Ringtone*, are well above twice the average (shown on the far right). Three categories – *Money*, *Cars*, and *Furniture* – are well below half the average. We also calculated *DCG* (*Discounted Cumulated Gain*) [10] spam densities, which give more weights to spam URLs appearing near the top of the search-result list, but found no significant difference from Figure 2.

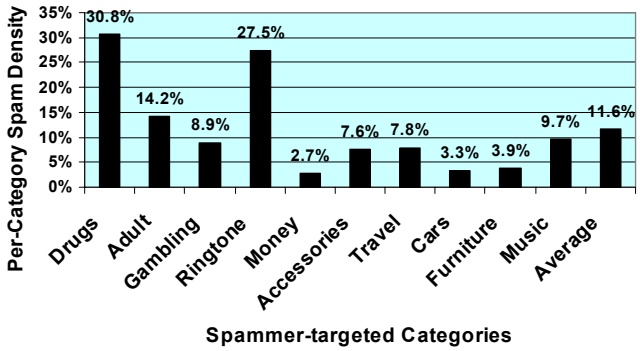

Figure 2: Per-category and average redirection-spam densities

4.2 Double-Funnel Analysis

We now analyze the five layers of the double-funnel, identify major domains involved at each layer, and categorize them to provide insights into the current trends of search spamming.

4.2.1 Layer #1: Doorway Domains

Figure 3 illustrates the top-15 primary domains/hosts by the occurrences of doorway URLs hosted on them. The first one is blogspot.com, with ***3,882*** appearances (of 2,244 unique doorway URLs), which is ***an order of magnitude higher*** than the others in the chart. This translates into a ***2.6%*** spam density by blogspot URLs alone, which is around ***22%*** of all detected spam appearances. (By comparison, the last one in the chart blog.hlx.com has 110 occurrences of 61 unique URLs.) Typically, spammers create spam blogs, such as http://PhentermineNoPrescriptionn.blogspot.com, and use these doorway URLs to spam the comment area of other forums. Since #2, #3, #4, and #7 in Figure 3 all belong to the same company, an alternative analysis would be to combine their numbers, resulting in ***1,403*** occurrences (***0.9%*** density) of 948 unique URLs.

The top-15 domains can be divided into four categories: five are **free blog/forum hosting sites**, five are **free web-hosting sites in English**, three appear to be **free web-hosting sites in foreign languages**, and the remaining two (oas.org and usaid.gov) are ***Universal Redirectors***, *which take an arbitrary URL as an argument and redirect the browser to that URL* [15]. For example, the known-spammer domain paysefeed.net, which appears to be exploiting tens of universal redirectors, was behind the following spam URLs: **http://www.oas.org/main/main.asp?slang=s&slink=**http://dir.kzn.ru/hydrocodone/ and **http://www.usaid.gov/cgi-bin/goodbye?**http://catalog-online.kzn.ru/free/verizon-ringtones/. We note that none of these 15 sites hosts only spam and therefore cannot simply be blacklisted by search engines. This confirms the anecdotal evidence that a significant portion of the web spam industry has moved towards setting up *"throw-away" doorway pages on legitimate domains*, which then redirect to their behind-the-scenes redirection domains, to be discussed in the next subsection.

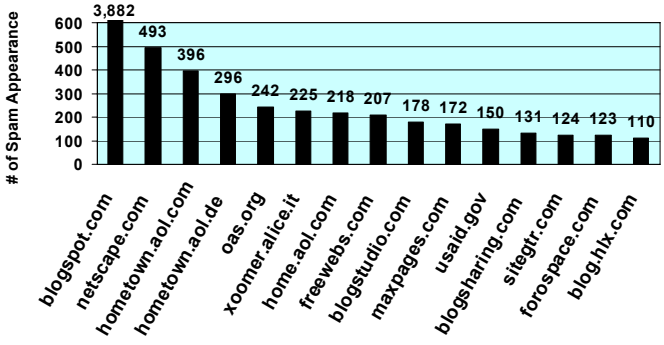

Figure 3: Layer #1: top-15 primary domains/sites by spam doorway appearance counts

Figure 3 is useful for search engines to identify spam-heavy sites to scrutinize their URLs. Figure 4 shows that 14 of the top-15 doorway domains have a ***spam percentage***[3] higher than ***74%***; that is, *3 out of 4 unique URLs on these domains (that appeared in our search results) were detected as spam*. To demonstrate the need for scrutinizing these sites, we scanned the top-1000 results from two queries – "*site:blogspot.com phentermine*" and "*site:hometown.aol.com ringtone*" – and identified more than half of the URLs as spam easily. It is in the interest of the owners of these legitimate websites to clean the heavy spam on their sites to avoid the reputation of spam magnets. We note that not all large, well-established web hosting sites are heavily abused by spammers. For example, in our data, each of tripod.com (#19), geocities.com (#32), and angelfire.com (#38) had fewer spam appearances than some newer, smaller web sites that rank among the top 15 in Figure 3.

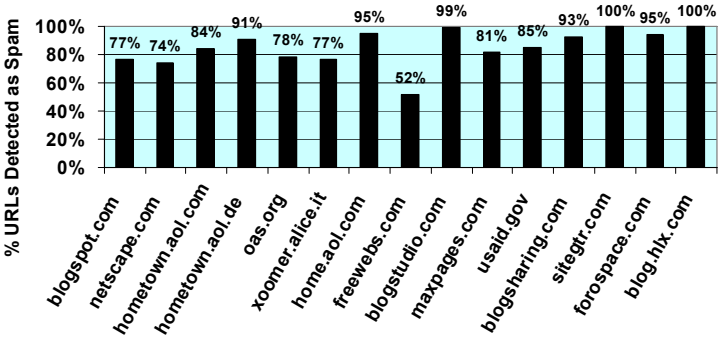

Figure 4: Layer #1: top doorway domains and their spam percentages (among the search results in our data)

Spam Pages on .gov and .edu Domains

When a site within a non-commercial top-level domain, such as .gov and .edu, occurs prominently in the search results of spammer-targeted commercial search terms, it often indicates that the site has been spammed. Figure 5 illustrates the 15 .gov/.edu domains that host the largest number of spam URLs in our data. These URLs can be divided into three categories:

[3] We note that "spam percentage" is calculated on a per-domain basis and is defined as the number of unique spam URLs divided by the number of unique URLs on a given domain that appeared in our search results.

(1) **Universal redirectors:** for example, these two spam URLs **http://serifos.eecs.harvard.edu/proxy/**http://catalog-online.kzn.ru/christian-ringtones/[4] and **http://www.fmcsa.dot.gov/redirect.asp?page=**http://maxpages.com/troctrocbas both redirect to paysefeed.net.

(2) **Unprotected upload areas,** such as http://uenics.evansville.edu:8888/school/**uploads**/1/buy-carisoprodol-cheap.html and http://xdesign.ucsd.edu/**twiki**/bin/view/main/tramadolonline.

(3) **Home page-like directories,** such as http://aquatica.mit.edu/albums/**gtin**/texas-country-ringtones.html and http://find.uchicago.edu/**~loh**/albums/ cial.php?id=56.

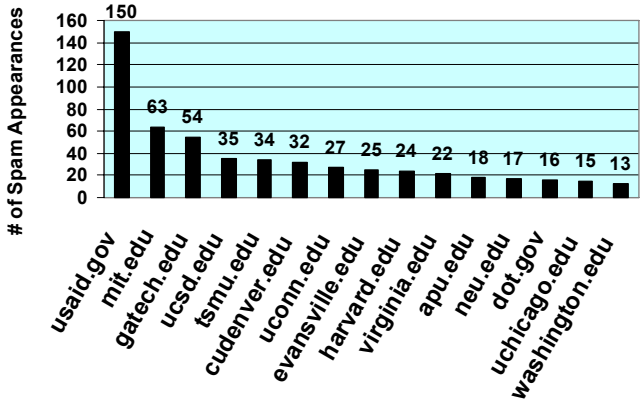

Figure 5: Layer #1: top-15 .gov/.edu domains

We observed that owners of the two domains nudai.com and raph.us appeared to be targeting .edu domains and were behind spam URLs hosted on 8 of the 15 domains. Another two ubiquitous spammers, paysefeed.net and topmeds10.com, covered 6 of the remaining 7 domains.

4.2.2 Layer #2: Redirection Domains

Figure 6 shows the top-15 redirection domains ranked by the number of spam doorway URLs that redirected to them. ***Twelve*** of them were syndication-based, serving text-based ads-portal pages containing 5 to 20 ads each, two of them displayed pornographic ads, and the remaining one was a commerce website. Domains #1, #2, #3, #5, and #10 all resided on the same IP block ***between 209.8.25.150 and 209.8.25.159***, collectively responsible for serving ads on ***3,909*** spam appearances (or ***2.6%*** spam density and ***22%*** of all detected spam appearances). Furthermore, topsearch10.com and searchadv.com shared the same registrant, and topmeds10.com and topmobile10.com shared the same proxy registrant. In addition, paysefeed.net and arearate.com shared the same registrant, while vip-online-search.info and webresourses.info shared the same IP address 195.225.177.32. In summary, a few major spammer groups appeared to own multiple top redirection domains.

None of the AdSense spammers appeared in the top-15 list. The highest-ranking AdSense spammer was ca-pub-4084532739617626 (#45), with 112 spam appearances of randomly named, made-for-ads .info domain pages, such as http://583.8d1w.info and http://101.j5bpqexcfs.info.

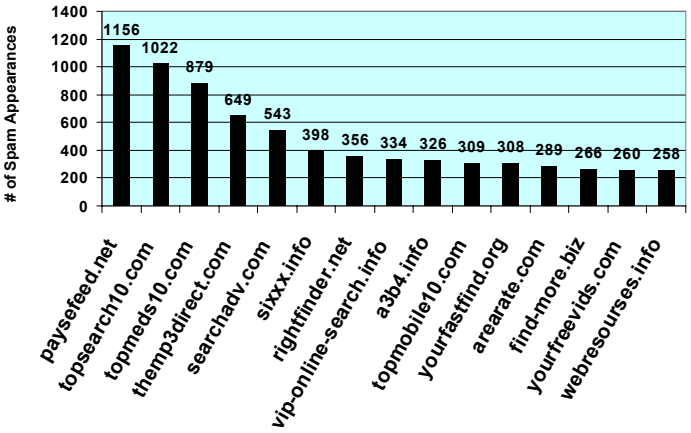

Figure 6: Layer #2: top-15 redirection domains by doorway-URL appearance counts

4.2.3 The Bottom Three Layers

Next, we focus on redirection spam pages that are ads portals. Among the 12,635 unique spam URLs, we extracted ***5,172*** ads-portal pages that contained a total of ***72,239*** ads and performed two types of analysis. For Layers #3 and #5, we performed ***page analysis*** by extracting target advertiser URLs as well as their associated click-through URLs from ads-portal pages, without visiting all the ads. For Layer #4, we performed ***click-through analysis*** by randomly selecting and visiting one ad from each portal page and recording all resulting redirection traffic. This was necessary because the domain names of intermediate syndicators did not appear in the content of ads-portal pages.

Layer #3: Aggregators (Page analysis)

Figure 7 illustrates the top-15 click-through traffic receiver domains based on analyzing static ads appearances on spam pages. Interestingly, all of them are in the form of IP addresses that can be divided into two groups: 13 of the IP addresses belong to the block ***between 66.230.128.0 and 66.230.191.255*** [30], which will be referred to as *"the 66.230 IP block"* throughout the paper, while the remaining two (#1 and #12) belong to the block ***between 64.111.192.0 and 64.111.223.255*** [30], to be referred to as *"the 64.111 IP block"*. We note that the two IP blocks actually share the same network Whois record.

In total, we collected ***51,392*** and ***8,186*** ads appearances for the 66.230 block and the 64.111 block, respectively. Furthermore, even for some of the ads with non-IP domain names, such as it-psp.com (#18) and abosearch.com (#19), their click-through traffic eventually still got funneled into the above two IP blocks. This suggests that if we had performed a more comprehensive click-through analysis of all the ads, we would have found even more ads-portal pages sending click-through traffic to these two IP blocks.

Layer #5: Advertisers (Page analysis)

On most spam ads, the click-through URLs did not contain the plaintext URLs of their target advertisers[5]. But the advertisers'

[4] We have notified several of the website owners, so the spam URLs reported in this paper may no longer be active by the time the paper is published. But one may still use the *"link:"* query to see where these URLs were comment-spammed.

[5] The click-through URLs did contain encoded URLs of the advertisers; however, decoding these URLs seemed non-trivial.

domain names were often displayed either as anchor text or in the status bar upon mouse-over. By extracting such domain names from the ads-portal pages and ranking them based on the number of their appearances, we plot in Figure 8 the top-15 advertisers (for the 10 categories that we studied): 10 are ringtone-related, two belong to the drugs category, one belongs to the money category, and the remaining two are cross-category. Well-known names that appeared on the complete list include: shopping.com (#22, 492), dealtime.com (#25, 465), bizrate.com (#33, 305), orbitz.com (#44, 258), ebay.com (#52, 225), and shopzilla.com (#54, 221).

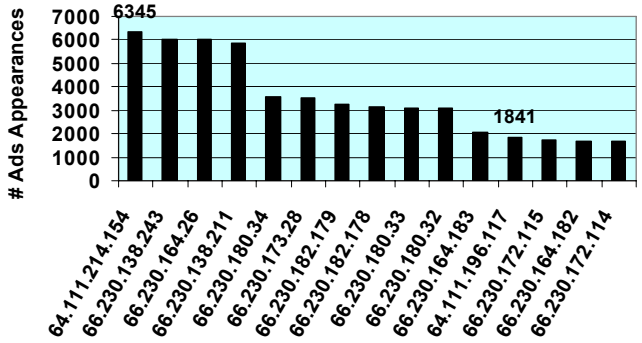

Figure 7: Layer #3: top-15 click-through traffic receiver domains by the number of ads appearances on spam pages (page analysis); the two numbers mark the only two IP addresses that belong to the 64.111 IP block.

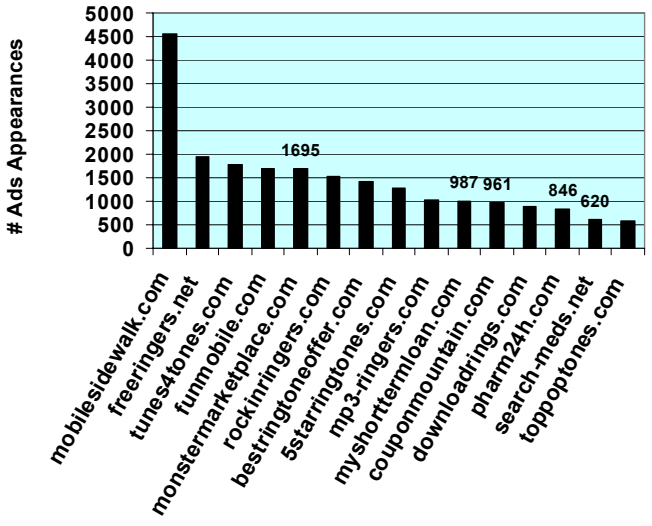

Figure 8: Layer #5: top-15 advertisers by the number of ads appearances on spam pages (page analysis); the five numbers mark the five non-ringtone advertisers.

Layer #4: Syndicators (Click-through analysis)

In the click-through analysis, a handful of syndicator domains had significant presence in the redirection chains. They appear to be the major middlemen between spam-traffic aggregators and the advertisers. In particular, the top-3 syndicators: findwhat.com, looksmart.com, and 7search.com appeared on 1,656, 803, and 606 redirection chains, respectively. (See [32] for sample screenshots.) They together accounted for 3,065 (***59%***) of the 5,172 redirection chains.

5. ADVERTISER-TARGETED KEYWORDS

In Section 4, we analyzed five layers of end-to-end search spam based on the most spammed keywords at public forums. However, the primary concern of most search users and legitimate advertisers is the impact of such spam on the quality of their query results. For example, they may not care if large amount of spam targets search terms outside their interest, such as online drug purchases. To answer this question, we repeat the analyses using a different benchmark based on the most-bid keywords from legitimate advertisers.

5.1 Benchmark of 1,000 Most-Spammed Advertiser-Targeted Keywords

For our second benchmark, we obtained a list of 5,000 most-bid keywords from a legitimate ads syndication program, queried them at all three major search engines to retrieve the top-50 results in early October 2006, scanned and analyzed all URLs with Search Ranger, and selected the 1,000 keywords with the highest per-keyword spam densities. Compared to the spammer-targeted benchmark in Section 3, this benchmark has fewer keywords from the drugs, adult, and gambling categories, and more keywords from the money category and other miscellaneous categories[6]. The two benchmarks overlap by 15%.

5.2 Spam Density Analysis

Overall, we scanned 95,753 unique URLs and identified ***6,153*** of them as spam, which accounted for ***5.8%*** of all top-50 appearances. This number is lower than the 11.6% number for the previous benchmark, and there are two partial explanations. First, this second benchmark has fewer keywords from the heavily spammed categories in Figure 2. Second, we measured the second benchmark two weeks after we measured the first one, while one of the three major search engines started to remove spam URLs right after our first measurement.

5.3 Double-Funnel Analysis

We next analyze the five layers and compare them with the results from the first benchmark. In all the figures, we color those domains that have appeared previously gray.

5.3.1 Layer #1: Doorway Domains

Figure 9 illustrates the top-15 doorway domains, five of which also appeared in Figure 3 and two were previously discussed .edu domains. Similar to Figure 3 and Figure 4, blogspot.com remained No. 1 with an-order-of-magnitude higher spam appearances than the other domains, accounted for ***29%*** of all detected spam appearances, and had a spam percentage as high as ***75%***. Again, all but one of the top-15 domains (uconn.edu in this case) had a higher than ***74%*** spam percentages (details omitted). The most notable differences from Figure 3 are the four .info domains, all of which appeared to have been set up solely for hosting doorway pages. In fact, 1,224 of the 1,798 unique .info URLs were identified as spam, and they had 1,324 appearances, ***15%*** of all detected spam. Table 1 shows that .info had a ***68%***

[6] Also we did not consider category information when determining this benchmark.

spam percentage in our search results, which is ***an-order-of-magnitude higher*** than that for .com (4.1%). (The two numbers were 63% and 9.6% for the spammer-targeted benchmark.)

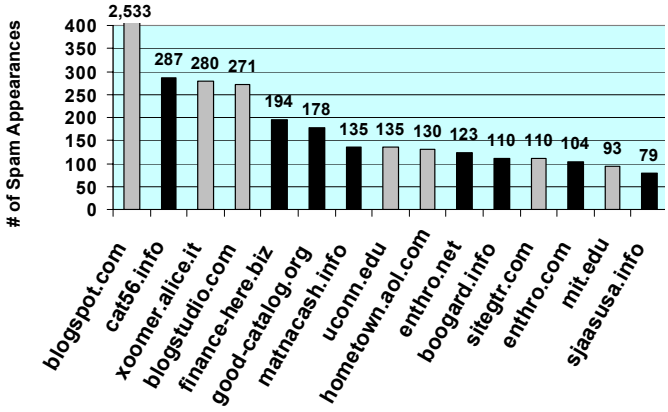

Figure 9: Layer #1: top-15 primary domains/sites by spam doorway appearance counts

TLD	.com	.org	.net	.biz	.info
Spam %	4.1%	11%	12%	53%	68%

Table 1: Spam percentages for Top-Level Domains (TLDs) based on search results in our second benchmark

5.3.2 Layer #2: Redirection Domains

Figure 10 shows the top-15 redirection domains, ***all of which were syndication-based***. Seven of them overlap with the list in Figure 6, and nudai.com was previously discussed. Topsearch10.com stands out as the only redirection domain that was behind over 1,000 spam appearances in both benchmarks. In addition, redirection domains residing in the ***209.8.25.150~209.8.25.159*** IP block continued to have a significant presence with 2,208 doorway appearances, which accounted for ***25%*** of all spam appearances. The most notable differences are that drugs and adult spammers are replaced by money spammers, reflecting the different compositions of the two benchmarks. Finally, we note that veryfastsearch.com (64.111.196.122) and nudai.com (64.111.199.189) belonged to the 64.111 IP block described in Section 4.2.3, and could potentially connect to the aggregator more directly. Again, none of the AdSense spammers appeared in the top-15 list. The highest-ranking one was ca-pub-2706172671153345, who ranked #31 with 61 spam appearances of 27 unique spam blogs at blogspot.com.

5.3.3 The Bottom Three Layers

Among the 6,153 unique spam URLs, we extracted ***2,995*** ads-portal pages that contained a total of ***37,962*** ads.

Layer #3: Aggregators (Page analysis)

Figure 11 shows that, again, the 66.230 and 64.111 IP blocks contained dominating receiver domains for spam-ads click-through traffic. In total, we collected ***28,938*** and ***6,041*** ads for these two IP blocks, respectively.

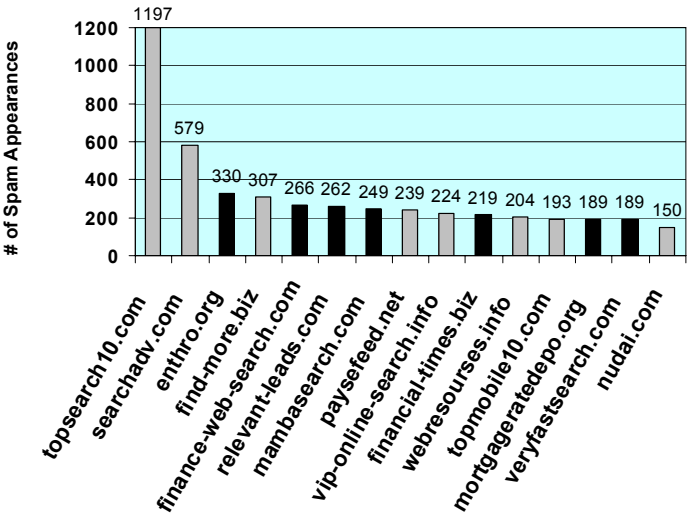

Figure 10: Layer #2: top-15 redirection domains by number of spam doorway appearances

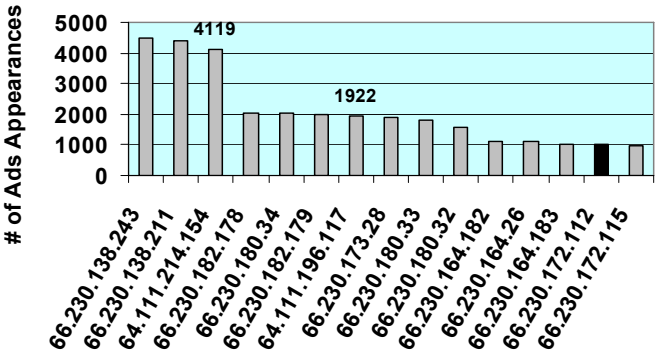

Figure 11: Layer #3: top-15 click-through traffic receiver domains by the number of ads appearances on spam pages (page analysis)

Layer #5: Advertisers (Page analysis)

Figure 12 identifies the top-15 advertisers, which are significantly different from the ones in Figure 8; only six of them overlap. Well-known sites – such as bizrate.com, shopping.com, dealtime.com, and shopzilla.com, which previously ranked between #20 and #60 – now move into the top 15. This reflects the fact that advertiser-targeted keywords better match these shopping websites than spammer-targeted keywords.

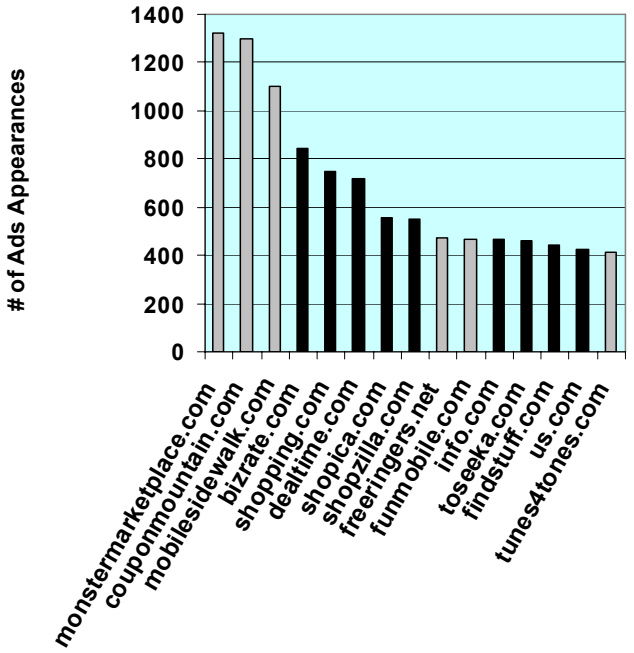

Figure 12: Layer #5: top-15 advertisers by number of ads appearances on spam pages (page analysis)

Layer #4: Syndicators (Click-through analysis)

Our click-through analysis shows that the two benchmarks shared the same list of top-3 syndicators, despite the fact that the benchmarks had only 15% overlap in the list of keywords and very different top-advertisers list. Again, the top-3 syndicators appeared on a large number of redirection chains in our analysis: looksmart.com (881), findwhat.com (809), and 7search.com (335), which together accounted for 2,025 (***68%***) of the 2,995 chains. These numbers demonstrate that these syndicators appear to be involved in the search spam industry both broadly and deeply.

6. OTHER COMMON SPAM

In this section, we show that many syndication-based spammers who do not use client-side browser redirections to fetch ads share the same bottom half of the double-funnel with redirection spammers; that is, although they fetch ads on the server side, they also funnel the click-through traffic from their pages into the same IP blocks that we uncovered in the previous sections. This shows that the aggregators and the syndicators are profiting from even more spam traffic. All scans were performed in the month of October 2006.

6.1 BLOG FARMS

The web page at http://urch.ogymy.info/ is a commonly seen made-for-ads blog page that consists of three parts: a list of ads, followed by a few programmatically generated short comments, followed by a long list of meaningless paragraphs designed to promote several randomly named .org and .info URLs sprinkled throughout the paragraphs. By issuing the following queries – *"Welcome to my blog" "Hello, thanx for tips" phentermine domain:info*, as well as *"linkdomain:ogymy.info"* and *"linkfromdomain:ogymy.info"* – we found 1,705 unique pages that shared the same format and belonged to the same blog farm. By visiting each page and analyzing the ads URLs, we found that all ***17,050*** ads forwarded click-through traffic to 64.111.196.117, which was #12 in Figure 7 and #7 in Figure 11.

6.2 PARASITE ADS-PORTAL FARMS

The web pages at http://phentermine.IEEEpcs.org/, http://www.HistMed.org/Gambling-Online.phtml, and http://ChildrensMuseumOfOakridge.org/PornStar-Finder.dhtml [32] are three examples of commonly seen made-for-ads pages that attach themselves to legitimate domains to increase their search ranking and to resist blacklisting. By searching for other farms with similar signatures, we discovered 91 .org domains that have been infected with such "parasites": 10 had been removed, 3 appeared as "Under Construction", and the rest were actively serving ads. By visiting 10 pages on each of the active farms, we extracted ***15,580*** ads and found that ***6,200*** of them were funneling click-through traffic to 64.111.210.10, 64.111.210.206, and 64.111.214.154 (#1 in Figure 7), all of which belong to the 64.111 IP block. The remaining ***9,380*** ads belong to 66.230.138.243 and 66.230.138.211, #2 and #4 in Figure 7, respectively. We observed that a few of the .org domains used click-through cloaking [22]; for example, http://www.urbanacademy.org/pc-fix-it.phtml returned "HTTP 404 Not Found" when visited directly, but displayed a page of ads when visited through a search-result click-through.

7. RELATED WORK

Cloaking and redirection are two techniques that Gyongyi and Garcia-Molina identified as tactics for hiding spam content [8]. Wu and Davison studied cloaking and redirection on the web and found that more than 8% of the top 200 URLs returned by Google employed cloaking and that some sites even used redirection cloaking, i.e., redirecting different user agents to different sites [23]. They proposed an automated method to detect semantic cloaking, which first identifies suspect pages by the content of the pages returned to a browser and a crawler, and then uses machine learning to create a classifier [25]. Our Search Monkeys are able to foil cloaking, including the newer click-through cloaking techniques, by mimicking search users' behavior using a full-fledged browser so that redirection analyses are performed on true pages displayed to the users.

Money is a major incentive for spammers. Jansen observed that despite the problem of click-fraud, sponsored search could reduce the amount of spam [9]. Sarukkai proposed a way to quantify a search term's monetizability [17]. Chellapilla and Chickering investigated cloaking from an economic perspective by comparing search results from the top 5000 queries and the top 5000 monetizable queries. They observed that for queries whose results used cloaking, 73.1% pages of the popular queries were spam while 98.5% pages of the monetizable queries were spam [5]. We focus on detecting large-scale spammers by following the money to track down major domains that appear in the redirection chains involving spam ads.

Various ranking mechanisms, such as Pagerank, HITS, and Trust Rank, incorporate the idea that a link is a "vote" of trust [13, 26]. Baeza-Yates, Castillo, and Lopez found that Pagerank was vulnerable to Sybil attacks in which pages with low score formed a complete subgraph or a star [2]. However, Adali et al argued that maximizing rank could be as simple as a link bomb consisting of one central page to which every other page links [1]. Methods for adapting ranking algorithms to combat link farms include investigating trust starting with a known-bad seed or

introducing a measure of distrust [26]. Krishnan and Raj used this idea for Anti-Trust Rank, in which they used an algorithm similar to Trust Rank to propagate anti-trust from an initial seed set [12], similar to the work for identifying "neighborhoods" of distrust on the web [13] and link farms [24]. Benczur and Csalognany presented Spamrank as an automated spam detection technique by identifying pages that violated the power law distribution by linking to one another [4]. They observed that link similarity measures could be more effective than trust/distrust measures in classifying spam pages. Similarly, Carvalho et al. focused on identifying "noisy" links, which are sites with abnormal support between each other, by measuring the amount of linking between two sites [6]. Becchetti et al analyzed the heuristics – purely link-based analyses, Pagerank, Trustrank, Truncated PageRank, and various combinations of these heuristics – for spam detection and compared their performance [3]. In contrast, we use link analyses only to identify spammed forums, but rely on redirection analysis to identify spam pages.

Content analysis is also useful for detecting spam. Kolari, Finn, and Joshi took a machine learning approach by building a classifier based on meta tags, anchor text, and tokenized URLs [11]. Fetterly, Manasse, and Ntoulas began with content independent heuristics, such as URL structure and average change throughout a site [7], and continued with site-dependent heuristics, such as the words used in a page or title and the fraction of visible content [16]. Urvoy et al modelized the style of HTML documents based on properties such as spacing and HTML tags to determine stylistic similarities that could be used to identify authors [18]. Mishne, Carmel, and Lempel compared the language model between a sample blog entry and the target page specified by a comment [14]. Our traffic-based analysis is complementary to these content-based analyses.

8. CONCLUSIONS

We have presented redirection-spam analyses using the Strider Search Ranger system, which detects spam pages by monitoring their redirection traffic to known-spammer domains. Using a benchmark of spammer-targeted keywords, we showed that "drugs" and "ringtone" were the two most-spammed categories with an average search-result spam density as high as 30.8% and 27.5%, respectively. We have also constructed a second benchmark of advertiser-targeted keywords in order to study the similar and different spam characteristics between the two benchmarks.

We have presented a five-layer double-funnel model for analyzing redirection spam, in which ads from merchant advertisers are funneled through a number of syndicators, aggregators, and redirection domains to get displayed on spam doorway pages, whereas click-through traffic from these spam ads is funneled, in the reverse direction, through the aggregators and syndicators to reach the advertisers. Domains in the middle layers provide the critical infrastructure for converting spam traffic to money, but they have mostly been hiding behind the scenes. We used systematic and quantitative traffic-analysis techniques to identify the major players and to reveal their broad and deep involvement in the end-to-end spam activities.

For Layer #1 – doorway domains, we showed that the free blog-hosting site ***blogspot.com*** had an-order-of-magnitude higher spam appearances in top search results than other hosting domains in both benchmarks, and was responsible for about ***one in every four*** spam appearances (22% and 29% in the two benchmarks respectively, to be exact). In addition, at least ***three in every four*** unique blogspot URLs that appeared in top-50 results for commercial queries were spam (77% and 75%). We also showed that over ***60%*** of unique .info URLs in our search results were spam, which was an-order-of-magnitude higher than the spam percentage number for .com URLs.

For Layer #2 – redirection domains, we showed that the spammer domain ***topsearch10.com*** was behind over 1,000 spam appearances in both benchmarks, and the ***209.8.25.150~209.8.25.159*** IP block where it resided hosted multiple major redirection domains that collectively were responsible for ***22-25%*** of all spam appearances. We also observed that the majority of the top redirection domains were syndication-based, serving text-based ads-portal pages.

For Layer #3 – aggregators, we presented the surprising finding that two IP blocks ***66.230.128.0~66.230.191.255*** and ***64.111.192.0~64.111.223.255*** appeared to be responsible for funneling an overwhelmingly large percentage of spam-ads click-through traffic. In our study, we easily collected over 100,000 spam ads that were associated with these two IP blocks, including many ads served by non-redirection spammers as well. These two IP blocks occupy the "bottleneck" of the spam double-funnel and may prove to be the best layer for attacking the search spam problem.

For Layer #4 – syndicators, we discovered that a handful of ads syndicators appeared to serve as the middlemen for connecting advertisers with the majority of the spammers. In particular, the top-3 syndicators were involved in ***59-68%*** of the spam-ads click-through redirection chains that we sampled. By serving ads on a large number of low-quality spam pages at potentially lower prices, these syndicators could become major competitors to main-stream advertising companies who serve some of the same advertisers' ads on search-result pages and other high-quality, non-spam pages.

For Layer #5 – advertisers, we showed that even well-known websites' ads had significance presence on spam pages. Ultimately, it is advertisers' money that is funding the search spam industry, which is increasingly cluttering the web with low-quality content and reducing web users' productivity. By exposing the end-to-end search spamming activities, we hope to educate users not to click spam links and spam ads, and to encourage advertisers to scrutinize those syndicators and traffic affiliates who are profiting from spam traffic at the expense of the long-term health of the web.

9. REFERENCES

[1] Adali, S., Liu, T., and Magdon-Ismail, M. Optimal Link Bombs are Uncoordinated. In *the 1st International Workshop on Adversarial Information Retrieval on the Web (AIRWeb),* May 2005.

[2] Baeza-Yates, R, Castillo, C., and Lopez, V. Pagerank Increase Under Different Collusion Topologies. In *the 1st International Workshop on Adversarial Information Retrieval on the Web (AIRWeb),* May 2005.

[3] Becchetti, L., Castillo, C., Donato, D., Leonardi, S., Baeza-Yates, R. Link-based Characterization and Detection of Web Spam. In *the 2nd International Workshop on Adversarial Information Retrieval on the Web (AIRWeb),* August 2006.

[4] Benczur, A., Csalogany, K., Sarlos, T., and Uher, M. SpamRank – Fully Automatic Link Spam Detection. In *the*

1st International Workshop on Adversarial Information Retrieval on the Web (AIRWeb), May 2005.

[5] Chellapilla, K. and Chickering, D.M. Improving Cloaking Detection Using Search Query Popularity and Monetizability. In *the 2nd International Workshop on Adversarial Information Retrieval on the Web (AIRWeb),* August 2006.

[6] da Costa Carvalho, A. L., Chirita, P., de Moura, E. S., Calado, P., and Nejdl, W. Site Level Noise Removal for Search Engines. In *Proc. of International World Wide Web Conference (WWW).* May, 2006.

[7] Fetterly, D., Manasse, M., and Najork, M. Spam, Damn Spam, and Statistics: Using Statistical Analysis to Locate Spam Web Pages. In *Proc of the 7th International Workshop on the Web and Databases.* pp. 1-6, 2004.

[8] Gyongyi, Z. and Garcia-Molina, H. Web Spam Taxonomy. In *the 1st International Workshop on Adversarial Information Retrieval on the Web (AIRWeb)*, 2005.

[9] Jansen, B.J. Adversarial Informaton Retrieval Aspects of Sponsored Search. In *the 2nd International Workshop on Adversarial Information Retrieval on the Web (AIRWeb),* 2006.

[10] Jarvelin, K. and Kekalainen, J. IR Evaluation Methods for Retrieving Highly Relevant Documents. In *Proc. ACM SIGIR Conference on R&D in Information Retrieval*, 2000

[11] Kolari, P., Finin, T., and Joshi, A. SVMs for the Blogosphere: Blog Identification and Splog Detection. In *AAAI Spring Symposium on Computational Approaches to Analysing Weblogs*, March 2006.

[12] Krishnan, V. and Raj, R. Web Spam Detection and Anti-Trust Rank. In *the 2nd International Workshop on Adversarial Information Retrieval on the Web (AIRWeb),* August 2006.

[13] Metaxas, P. and DeStephano, J. Web Spam, Propaganda and Trust. In *the 1st International Workshop on Adversarial Information Retrieval on the Web (AIRWeb),* May 2005.

[14] Mishne, G., Carmel, D., and Lempel, R. Blocking Blog Spam with Language Model Disagreement. In *the 1st International Workshop on Adversarial Information Retrieval on the Web (AIRWeb),* May 2005.

[15] Niu, Y., Wang, Y. M., Chen, H., Ma, M., and Hsu, F. A Quantitative Study of Forum Spamming Using Context-based Analysis. In *Proc. Network and Distributed System Security (NDSS) Symposium*, February 2007.

[16] Ntoulas, A., Najork, M., Manasse, M., and Fetterly, D. Detecting Spam Web Pages through Content Analysis. In *Proc. International World Wide Web Conference (WWW),* May 2006.

[17] Sarukkai, R.R. How Much is a Keyword Worth? In *Proc. International World Wide Web Conference, (WWW)*, May 2005.

[18] Urvoy, T., Lavernge, T., Filoche, P. Tracking Web Spam with Hidden Style Similarity. In *the 2nd International Workshop on Adversarial Information Retrieval on the Web (AIRWeb),* August 2006.

[19] Wang, Y. M., Beck, D., Jiang, X., Roussev, R., Verbowski, C., Chen, S., and King, S. Automated Web Patrol with Strider HoneyMonkeys: Finding Web Sites That Exploit Browser Vulnerabilities. In *Proc. Network and Distributed System Security (NDSS) Symposium*, February 2006.

[20] Wang, Y. M., Beck, D., Wang, J., Verbowski, C., and Daniels, B. Strider Typo-Patrol: Discovery and Analysis of Systematic Typo-Squatting. In *Proc. 2nd Workshop on Steps to Reducing Unwanted Traffic on the Internet (SRUTI)*, July 2006.

[21] Wang, Y. M. and Ma, M. Strider Search Ranger: Towards an Autonomic Anti-Spam Search Engine. Microsoft Research Technical Report, MSR-TR- 2006-174, December 2006

[22] Wang, Y. M. and Ma, M. Detecting Stealth Web Pages That Use Click-Through Cloaking. Microsoft Research Technical Report, MSR-TR- 2006-178, December 2006

[23] Wu, B. and Davison, B.D. Cloaking and Redirection: A Preliminary Study. In *the 1st International Workshop on Adversarial Information Retrieval on the Web (AIRWeb),* 2005.

[24] Wu, B., and Davison, B.D. Identifying Link Farm Pages. In *Proc. International World Wide Web Conference (WWW)*, 2005.

[25] Wu, B. and Davison, B.D. Detecting Semantic Cloaking on the Web. In *Proc. International World Wide Web Conference (WWW),* August 2006.

[26] Wu, B., Goel, V., Davison, B.D. Propagating Trust and Distrust to Demote Web Spam. In *Proc. Models of Trust for the Web Workshop (MTW), International World Wide Web Conference*, 2006.

[27] Fiddler HTTP Proxy, http://www.fiddlertool.com/

[28] Fighting Splogs, http://fightsplog.blogspot.com/

[29] The Google AdSense Program, http://google.com/adsense

[30] Network Whois records, http://whois.domaintools.com/ 66.230.138.211 and http://whois.domaintools.com/ 64.111.214.154

[31] Screenshots of sample redirection spam pages, http://research.microsoft.com/SearchRanger/Redirection-spam_3_types.htm

[32] Screenshots of sample click-through analyses, http://research.microsoft.com/SearchRanger/Spam_ads_click-through_analysis.htm