A Quantitative Study of Forum Spamming Using Context-based Analysis

A Strider Search Ranger Report

Yuan Niu
Yi-Min Wang
Hao Chen
Ming Ma
Francis Hsu

December 1, 2006

Technical Report
MSR-TR-2006-173

Microsoft Research
Microsoft Corporation
One Microsoft Way
Redmond, WA 98052

To appear in *Proc. Network & Distributed System Security (NDSS) Symposium*, Feb. 2007.

A Quantitative Study of Forum Spamming Using Context-based Analysis

Yuan Niu†, Yi-Min Wang‡, Hao Chen†, Ming Ma‡, and Francis Hsu†

†University of California, Davis {niu,hchen,hsuf}@cs.ucdavis.edu
‡Microsoft Research, Redmond {ymwang,mingma}@microsoft.com

Abstract

Forum spamming has become a major means of search engine spamming. To evaluate the impact of forum spamming on search quality, we have conducted a comprehensive study from three perspectives: that of the search user, the spammer, and the forum hosting site. We examine spam blogs and spam comments in both legitimate and honey forums. Our study shows that forum spamming is a widespread problem. Spammed forums, powered by the most popular software, show up in the top 20 search results for all the 189 popular keywords. On two blog sites, more than half (75% and 54% respectively) of the blogs are spam, and even on a major and reputably well maintained blog site, 8.1% of the blogs are spam[1]. The observation on our honey forums confirms that spammers target abandoned pages and that most comment spam is meant to increase page rank rather than generate immediate traffic. We propose context-based analyses, consisting of redirection and cloaking analysis, to detect spam automatically and to overcome shortcomings of content-based analyses. Our study shows that these analyses are very effective in identifying spam pages.

1 Introduction

Search engine spamming (or *search spamming* or *web spamming* [1]) refers to the practice of using questionable search engine optimization techniques to improve the ranking of web pages in search results. When search spamming started, spammers created a large number of web pages with crafted keywords and link structures to promote the ranking of their web sites in search results. As search engines develop more sophisticated techniques to identify spam web pages, spammers are moving towards more fertile playgrounds: *web forums*. We define a *web forum* as a web page where visitors may contribute content. Web forums range from the older message boards and guestbooks to the newer blogs and wikis. Since forums are designed to facilitate collaborative content contribution, they become attractive targets for spammers. For example, spammers have created spam blogs (splogs) and have injected spam comments into legitimate forums (See Table A1 in Appendix for sample spammed forums). Compared to the "propagandist" methods of web spamming through the use of link farms[2] [16], forum spamming poses new challenges to search engines because (1) it is easy and often free to create forums and to post comments in processes that can often be automated, and (2) search engines cannot simply blacklist forums that contain spam comments because these forums may be legitimate and contain valuable information. Anecdotal evidence as well as our own experience indicates that spammers have successfully promoted their web sites in search results through forum spamming. To combat forum spamming and to gain insight on effective defense mechanisms, we need a comprehensive study of the problem, such as its scale and its popular techniques.

This paper reports our quantitative study of forum spamming. We examine the scale of forum spamming from three different perspectives: that of the search user, the spammer, and the forum hosting site. We examine both spam blogs, and spam comments in forums. For such a large-scale study, manual identification of spam would be unscalable, so we need automated approaches. Most existing automatic spam detection approaches are based on content analysis. However, since spammers have discovered many ways to circumvent content-based analysis (such as plagiarized legitimate content, redirection[3], and

[1] We detected only redirection spam, so the actual percentage of spam blogs could be much higher.

[2] A link farm consists of a group of sites using specific linking structures to boost rankings of one or more pages in the farm.

[3] Redirection changes the user's final destination or fetches dynamic content from other web sites.

cloaking[4]), we propose context-based analyses of redirection and cloaking.

1.1 Three perspectives of forum spamming

First, we examine forum spamming from the search user's perspective: when a user searches forums, how much spam shows up among the top results? To answer this question, we search for 190 top keywords in forums powered by nine most popular forum programs. We will describe our findings in Section 3.1.1.

Second, we examine forum spamming from the spammer's perspective: we try to understand the spammer's modus operandi. To this end, we set up three honey blogs, which should attract no legitimate comments, and have collected 41,100 comments over a year. We will analyze these spam comments in Section 3.1.2.

Finally, we examine forum spamming from the forum hosting site's perspective: how many forums are spam on a hosting site, and what typical techniques do spammers apply to avoid detection? We examine over 20,000 blogs from four blog sites and will describe our findings in Section 3.2 and 3.3.

1.2 Context-based analysis

To avoid exposing their spam domains directly and being blacklisted by search engines, many spammers are creating *doorway pages* on free-hosting domains and using their URLs in comment spamming. When a user clicks a doorway-page link in search results, her browser is instructed to either *redirect to* or *fetch ad listings from* the actual spam domain or advertising companies that serve the spammer as a customer. For example, the comment-spammed URL *http://mywebpage.netscape.com/fendiblog/fendi-handbag.html* redirects to the well-known domain *topsearch10.com*.

Many spammers set up doorway pages on free forum hosting websites such as *blogspot.com*, *blogstudio.com*, *forumsity.com*, etc. Such doorway pages are a form of spam blogs. For example, as of this writing, *http://seiko-diver-watch.blogspot.com* appeared as the top Yahoo! search result for "Seiko diver watch", and *http://forumsity.com/mobile/1/free-motorola-ringtone.html* appears as the top MSN search result for "free motorola ringtone". Figure 1 illustrates the end-to-end search spamming activity involving both spam blog creation and spam comment injection.

To overcome the limitations of content-based spam detection, we propose an orthogonal *context-based approach* that uses *URL-redirection* and *cloaking analysis*. Our work was primarily motivated by two key observations:

- Many spam pages use cloaking and redirection techniques [1,4] so that search engines see different content than human users. A common technique is to serve page content that the browser will dynamically rewrite through script executions but the crawler will not. Our approach is to treat each page as a dynamic program, and to use a "monkey program" [6] to visit each page with a full fledged browser (so that the program can be executed in full fidelity) while analyzing the redirection traffic.
- Many successful, large-scale spammers have created a huge number of doorway pages that generate redirection traffic to a single domain that serves the spam content. By identifying those domains that serve content to a large number of doorway pages, we can catch major spammers' domains together with all their doorway pages and doorway domains.

1.3 Contributions

We make the following contributions:

- We conduct a comprehensive, quantitative study of forum spamming. We examine this problem from three perspectives: that of the search user, the spammer, and the hosting site.
- To overcome the limitation of content-based spam detection, we propose context-based detection techniques that use URL redirection tracing and cloaking analysis. We show that our context-based detection is effective in identifying spam pages automatically. Particularly, the two domains of a prolific spammer, who created a large percentage of spam on two blog sites, appear in the top results returned by our analysis tool.
- Our study confirmed that forum spamming is a widespread problem. The observation on our honey forums confirms that spammers target abandoned pages and that most comment spam is meant to increase page rank rather than generate immediate traffic.

[4] Cloaking serves different content to different web visitors. For example, it may serve browsers a page with spam content while serving crawlers a page optimized for improving ranking.

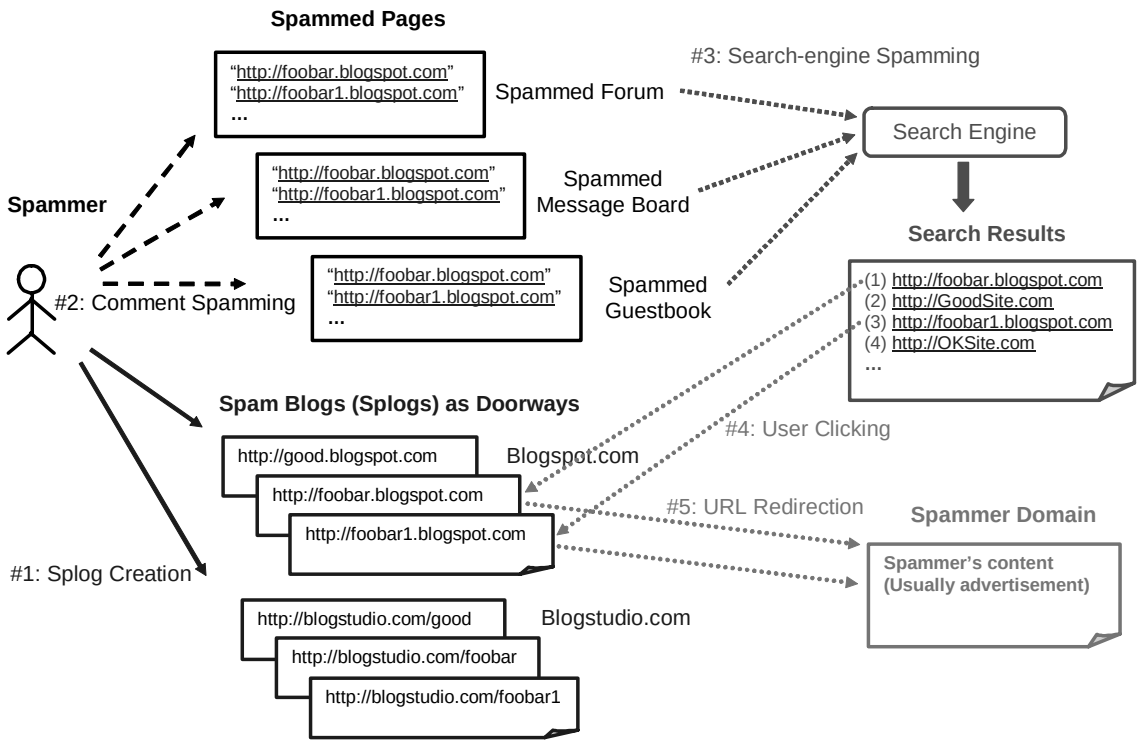


Figure 1: Spam Blogs (splogs) and Comment Spamming. #1: The spammer creates splogs. #2: The spammer spams forums with the URLs of his splogs. #3: These splog URLs are ranked high in search results. #4: The search user clicks the splog URLs in the search results. #5: The splogs redirect the user to the spammer domain.

2 Redirection-based spam detection

2.1 URL redirection tracing and analysis

Our context-based spam detection process starts with the collection of a list of "URLs of interest". Such URLs can be gathered from spammed forums, website and blog hosting sites, or search results, etc. We then feed the list of URLs to the *Strider URL Tracer* [5]. The tracer provides a key functionality called the *Top Domain view*: given a list of (primary) URLs, the tracer launches an actual browser to visit each URL and records all secondary URLs visited as a result. At the end of the batched scan, the Top Domain view provides the list of third-party domains that received secondary-URL traffic and rank them by the number of primary URLs that generated traffic to them. If the input is a list of highly suspicious spam URLs (such as those collected from a spammed forum), the Top Domain view highlights those behind-the-scenes *spammer domains* that are associated with a large number of doorway URLs.

If the input consists of a mix of spam and non-spam URLs, certain web sites, such as legitimate ads syndicators and web-analytics servers, may crowd the Top Domain view because they serve a large number of non-spam URLs. To filter out the noise, we scanned the top one million (mostly non-spam) click-through URLs [6] to obtain a "known-good whitelist" of top third-party domains and filtered them out from the Top Domain view. The remaining ranked Top Domain list is then used to prioritize manual investigation. Once a third-party domain is determined to be a spammer domain, it is added to the "known-bad blacklist" and will be excluded from future Top Domain view so that newer but smaller redirection receiver domains are exposed near the top. All doorway URLs associated with a blacklisted domain are labeled as high-potential spam URLs and will be referred to as *redirection spam*. How such information is utilized in a search ranking algorithm depends on each individual search engine. We note that our approach can be viewed as a combination of behavior-based and signature-based spam detection: Top Domain analysis is behavior-based, while third-party domain blacklisting is signature-based.

It is possible that some of the legitimate ads syndicators may have a significant number of spammers among their customers, in which case their domains are removed from the whitelist so that their

redirection traffic remains present in the Top Domain view. For those ads syndicators that embed their client IDs in the redirection URLs (e.g., *googlesyndication.com*), the tracer provides a second-level grouping and ranking of primary URLs by client IDs. Doorway URLs associated with a blacklisted client ID are also referred to as redirection spam in this paper.

2.2 Cloaking

To achieve effective search spamming and evade spam detection, spammers are increasingly using sophisticated cloaking techniques, which serve different content to different visitors. *Crawler-browser cloaking* is intended to fool the search engines into giving the spam URLs high ranks, while allowing spammers to serve ad pages that do not deserve the high ranks to the users. The cloaking server either provides different page content based on the User Agent field or known crawler IP addresses, or provide pages that contain scripts that will rewrite the page content (since most crawlers do not execute scripts, they will not see the rewritten content, while most browsers will execute the scripts to render the rewritten content to users).

Click-through cloaking, a new and lesser-known technique, attempts to fool human spam investigators and automatic spam-detection tools that directly visit the spam URLs instead of clicking through search results. It is primarily based on browser referrer checking and can be done in two ways: if the spammers own the website that hosts the spam URL, they can perform a server-side check of the Referer field in the HTTP header and serve different pages based on that. If the spam URL resides on a free hosting website, the spammer can serve a page containing a script that performs a client-side check of the browser document.referrer object and displays different pages based on that. For example, this spam URL http://gucci-handbag.bigcityhandbags.com/sondra-roberts-squared/song-titles-with-handbag/ was doing both checks: it used HTTP 302 to redirect to topsearch10.com if the HTTP request came with the right Referer field; otherwise, it served a client-side check script that used document.write() to display a bogus "Account is suspended" page if the check suggested that the visit did not come from a search click-through. (See Figure A1 in Appendix for more details.)

We have incorporated anti-cloaking techniques into the URL Tracer by making every visit from the tracer appear to come from a search click-through for both server-side and client-side *referer* checks. Furthermore, we use a diff-based technique to turn spammers' cloaking activities against themselves: for every suspicious URL that does not redirect to any known-bad domains, we scan it twice with anti-cloaking on and off, respectively, and take a diff of the resulting two URL redirection lists. If there is a significant discrepancy, the URL is highlighted for manual investigation. Once the URL is confirmed to be spam, its associated spam domain(s) are added to the blacklist and used by all future anti-cloaking scans to identify similar URLs associated with the same spammer.

3 Analysis of forum spam

3.1 Comment spam

3.1.1 Top forums and top search keywords. In this section, we examine forum spamming from the search user's perspective: how likely will the search user encounter spam in the forum search results? Since it is difficult to select a set of *top forums* objectively because forums cover very diverse topics, we selected *top forum software programs* instead. We chose nine of the most popular forum programs: *WWWBoard*, *Hypernews*, *Ikonboard*, *Ezboard*, *Bravenet*, *Invision Board*, *Phpbb*, *Phorum*, and *VBulletin*.

We created a list of 190 top keywords using popular tags from directories like *http://technorati.com*, *http://weblogs.com*, *http://metafilter.com*, and *http://icerocket.com* as well as lists of commonly spammed terms from *http://codex.wordpress.org*. Using both Google and MSN, we collected the top 20 results from searches for all the pairs of top keywords and top forum program names. To identify heavily spammed forums, we looked for sites that appeared in the search results for different keywords as well as sites whose pages appeared multiple times in the results for a single keyword.

Table 1: Most Heavily Spammed Forums

Forum	Pages	Keywords
http://fs.fed.us/cgi-bin/HyperNews_mm/get/mmforumA.html	175	102
http://www.comm.fsu.edu/interactive/forum/	134	82
http://www.usra.edu/phorum	119	94
http://classicauthors.net/messageboard/list.php?f=1	117	97
http://samba.eecs.umich.edu/phorum/list.php?2	105	79
http://phorum.nettwerk.com/	101	73

Table 1 shows the most heavily spammed forums, identified by the total number of keywords whose search results contain the forum's URL, and by the

total number of unique pages from the forum in the combined Google and MSN search results. Even forums on government sites were not immune from spamming, as shown in Table 2. In both Google and MSN search, we found spammed forums within the top 20 search results for each type of the top forum software for 189 of the 190 keywords (with "palm-texas-holdem-game" being the only exception). This shows that spam has significantly affected forum search users.

Table 2: Spammed Forums on Government Sites

Forum	Software
http://giscouncil.oshpd.ca.gov/viewforum.php?f=2&sid=9c8902ab2604e74f29f30eb1abf3d64b	PhpBB
http://www.deltastate.gov.ng/wwwboard/wwwboard.html	WWWBoard
http://www.hants.gov.uk/forum/ikonboard.cgi?s=8f5b05787d185c281f208403dd9dd79d;act=ST;f=12;t=504;&#top	Ikonboard
http://mdk.kinmen.gov.tw/phorum332c/	Phorum

3.1.2 Honey blogs. We set up three honey blogs powered by Wordpress on September 25, 2005 and posted at irregular intervals. The forum *litlog* contained quotations from literature (final post: December 7, 2005), *ilium* focused on mythology (final post: January 18, 2006), and *yabi* was an online diary which contained no useful information (final post: December 5, 2005). We configured each to be as open for commenting as possible, e.g., no moderation and no keyword blacklist. To attract spammers, we pinged *http://rpc.pingomatic.com* each time when we created a post so that our URL appeared on the list of recently updated blogs at various sites. We also linked the blogs to each other and provided one incoming link from a legitimate blog already indexed by several search engines. To date, the three honey blogs have received only two legitimate comments, among a total of 41,100 comments. We define a "post" as the content that the blog owner adds. "Comments" are any contributions made to the post through use of the comment form, trackbacks, or pingbacks. Trackbacks are pings sent to a user-specified URL when a blog post is published. The trackback appears on the comment page of the user specified URL as a comment that contains the URL of the blog post. Pingbacks are automated pings sent to every URL within a blog post. It too appears on the comment page of the URLs specified, as long as those sites also support pingbacks.

Temporal analysis. Figure 2 shows the number of accumulative comments received by each of the three honey forums during the first 339 days, from September 25, 2005 to August 30, 2006. The majority of the comments received by each honey blog arrived after we stopped posting content. For example, yabi,

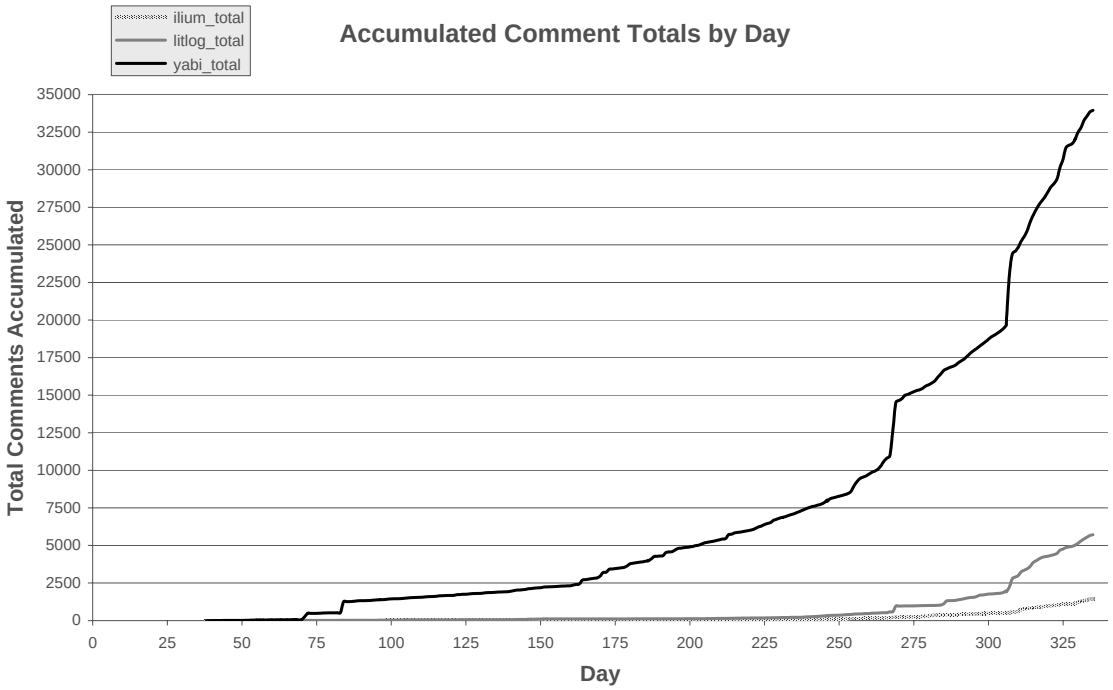


Figure 2: Accumulated comments received by each of three honey forums

the most popular of the three honey blogs, received 33,954 comments in its first 339 days but the majority of comments were received after March 2006 (starting from Day 188). On its busiest day (August 2, 2006, day 311), yabi received 3,142 new comments distributed over 21 posts. By August 30 (day 339), ilium amassed a total of 1,432 comments and litlog 5,714 comments. Ilium and litlog, which we updated less regularly, did not observe a dramatic increase in comments until well after 200 days after their creation. This supports the observation that spammers target abandoned pages and that most comment spam is meant to increase page rank, not to generate immediate traffic.

Figure 2 shows the rapid growth of comment totals for each blog. Yabi's growth was the most prolific, but all three clearly show that the bulk of comments were received at increasingly high rates in the last 30 days. In yabi's case, half of its total comments were received in the last 30 days, and the majority of comments for ilium and litlog came in this time period as well.

It is worth noting that the amount of spam continued to grow faster after the last day in Figure 2. In the 10 days between August 30 and September 9 (day 339 – 349), yabi received 3,449 new comments, almost 10% of the total comments received; ilium gained 305 new comments, 17.3% of its total; and litlog gained 1,081 comments, 16% of its total. The increased spam rate as these blogs got older could be because spamming had increased overall on the web or because spammers had identified these three blogs as "abandoned."

3.2 Spam blogs and cloaking analysis

In this section, we analyze sample blogs from three different blog sites --- *blogspot.com*, *blogspoint.com*, and *blogstudio.com* --- to evaluate the prevalence of splog doorways and cloaking. The numbers for spam that we report in the following are lower bounds because we only investigated spammer domains that received secondary traffic from a significant number of primary URLs (see the discussion on Top Domain view in Section 2.1).

3.2.1 Blogspot analysis. *Blogspot.com*, or *eBlogger*, is one of the most popular blog hosting sites. We randomly sampled 1% of the 1,749,150 blogspot profile pages created in July of 2006 that contained at least one blog link, and extracted all the blog links from these profile pages. In this way, we collected 19,271 blogspot URLs, of which 5,882 (31%) had been removed. Among the remaining 13,389 active blogs, our redirection analysis identified 1,091 (8.1%) pages as spam. Among the 1,091 spam blogs, 652 (60%) used click-through cloaking, and the top third-party domain *filldirect.com* was behind 259 (24%) cloaked splogs.

3.2.2 Blogspoint analysis. *Blogspoint.com* is a much smaller blog site. On August 23, 2006, we extracted and scanned its entire 4,714 blogs from their "memberlist" page. Our redirection analysis identified 3,535 (75%) pages as spam. Among the 3,535 spam pages, 2,166 redirected to *finance-web-search.com* and 917 redirected to *casino-web-search.com*. These top two third-party domains were registered by the same spammer with a Russian address, who most likely created these 3,083 splogs (65% of the entire 4,714 blogs and 91% of the 3,398 identified splogs). Because this spammer did not use cloaking, the cloaking percentage was much smaller than the blogspot number: only 3.9%, or 131 URLs.

Table 3: Number of Spam Blogs at Four Blog Hosting Sites

Blog Host	Examined URLs	Spam URLs	% of Spam	URLs using Cloaking
Blogspot	13,389	1,091	8.1%	652
Blogspoint	4,714	3,535	75%	131
Blogstudio	369	198	54%	0
Blogsharing	99	82	83%	0

3.2.3 Blogstudio analysis. *Blogstudio.com* is a lesser-known blog site, but its blog URLs appeared in a large number of spammed forums. Because some of these URLs had appeared among top query results at major search engines, we adopted the following sampling method: we issued a "*site:blogstudio.com*" query to MSN Search, retrieved the top 1000 results, and extracted 369 unique blogs. Our analysis identified 198 (54%) blogs as redirection spam. The same major blogspoint spammer owned 184 of these splogs (50% of all the sampled blogs and 93% of all the redirection spam), with 130 redirecting to *http://finance-web-search.com* and 54 redirecting to *http://casino-web-search.com*. None of these splogs used cloaking.

3.2.4 Summary. Table 3 summarizes the number of splogs at the above three blog hosts (the table also shows splogs at *Blogsharing*, which we will discuss in Section 3.3). Although only fewer than 10% of the sampled blogspot URLs were identified as redirection spam, the fact that we could find over a thousand splogs by sampling only 1% of the profiles created in one month suggests that the total number of redirection splogs could be in the millions. The surprisingly high percentage of splogs using cloaking (3 out of every 5) suggests that *blogspot.com* might have been actively identifying and removing less sophisticated spam, but

spammers who used click-through cloaking might have successfully evaded detection. This demonstrates the importance of using anti-cloaking techniques in spam detection and investigation.

Our analysis showed that smaller blog sites were heavily spammed by redirection spammers: in both *blogspoint* and *blogstudio*, more than half (75% and 54% respectively) of the analyzed blogs were redirection spam. A single spammer created an overwhelmingly large percentage (over 90%) of the splogs as doorways. His two spam domains were exposed prominently at the top of our Top Domain view, which convincingly demonstrates the effectiveness of our approach.

3.3 Second-level spam blogs

Blogsharing.com is another smaller blog site. Its spam blogs started appearing in top-10 search results in August 2006. Between August 23 and September 4, we collected 99 blogs from its "recently registered users" pages and our analysis identified 42 of them as redirection spam. In addition, we observed other tens of blogs that appeared to be created by the same spammer and shared the same format: the blog pages themselves were not doorways, but they contained spam URLs in the midst of junk text that redirected to the same spammer domain. For example, this blog *http://www.blogsharing.com/undangstrtresda/* contained the following three spam URLs: *http://plumbers-plumbing.com*, *http://the-plumber.info*, and *http://smartplumbers.info*, all of which redirected to *http://1rrk.com/plumber*. Essentially, these "*second-level spam blogs*" played the same role as a spammed forum: they provided links to spam URLs to improve their search ranking by exploiting link-counting algorithms that search engines use.

By first extracting links from each of the blogs and then performing redirection scan and analysis on those links, we identified another 40 spam blogs, each containing a few plumber-related links. In total, we found 58 unique plumber-related links, all of which redirected to *http://1rrk.com/plumber*. So the total number of splogs becomes 82, which account for *83%* of the 99 sampled blogs.

4 Other observations

4.1 Universal redirectors

We use the term *universal redirectors* to refer to legitimate websites' URLs that accept and redirect visiting browsers to arbitrary third-party URLs. A well-known universal redirector is *http://www.google.com/url?q=[any URL]*, which has been used by many phishers to make their phishing URLs look less suspicious; for example, this URL *http://www.google.com/url?q=http://213.190.10.80/chase.com/index.html* appeared in an actual phishing email targeting the Chase Manhattan Bank. Another universal redirector *http://rds.yahoo.com/_ylt=/*[any URL]* is also starting to get abused by email spammers.

During our manual investigation of comment spam, we noticed that a few universal redirectors hosted on university and government websites were used by some spammers as (a different form of) doorway URLs to redirect to spammer domains; examples included http://www.usaid.gov/cgi-bin/goodbye?http://catalog-online.kzn.ru/free/funny-ringtones/ (which appeared as the second result of a Yahoo query for "*funny ringtone*") and http://www.library.drexel.edu/cgi-bin/r.cgi?url=http://replica-watches.20six.co.uk. By searching through the forums where these URLs were comment-spammed, we were able to find 23 universal redirectors that were used by spammers, as shown in Table A2 in Appendix. As of August/September 2006, six of them had been disabled but the other 17 remained active.

Although many of these redirectors appear to have legitimate uses, the website owners should weigh the redirectors' benefits against their potential abuse. As hinted by the partially encoded spam URL found in the 9th universal redirector example, the spammer could have encoded the entire non-bold-face part of the URL, thus completely hiding the "spammy part" of the spam URL. (Search spamming could still be achieved through the anchor text not shown here.) Furthermore, as we will discuss in Section 4.2, many of the comment-spammed URLs are malicious, so an innocent universal redirector on a legitimate web site may serve inadvertently as an entry point through which malware spreads to many vulnerable machines.

4.2 Malicious spam URLs

A previous study [6] shows that a small percentage (0.071%) of the top one million URLs based on click-through counts at a search engine were malicious; they downloaded javascript code that attempted to exploit unpatched browser vulnerabilities on the visiting machines. A natural question to ask is whether malicious website operators are using comment spamming techniques to boost the ranking of their URLs. To answer this question, we investigated eight malicious URLs that appeared among the top-30 search results at the three major search engines between July and September 2006. For each URL, we used the search engine to try to find a forum that had been spammed with the URL. If we could find such a forum, we then extracted all URLs spammed on the same forum page and scanned them to see if we could identify additional malicious URLs. The results are shown in Table 4.

Table 4: Malicious URLs in Spammed Forums

	Malicious URLs	Spammed at (cached pages of) these forums	# Malicious URLs (%) of # Spam URLs
1	http://eclatlantus.yiffyhost.com/fessee-progenco.html	http://members.tripod.com/deportivolapigeon/guestbook.htm	688 (*47%*) of 1,463
2	http://humorrise.sitesled.com/burberry-handbag-supplier-wholesale.html	http://sgtrois.webator.net/?2004/11/02/4-un-cd-gratuit-offert-par-ladisq	6 (*18%*) of 33
3	http://alexandrsultz.sitesled.com/chloe/handbag.html	http://www.no-nation.org/article347.html	164 (*49%*) of 338
4	http://nowodus.tripod.com/new-mexico-lottery.html	http://aose.ift.ulaval.ca/modules.php?op=modload&name=News&file=article&sid=137	174 (*14%*) of 1,280
5	http://lermon.t35.com	http://68.15.204.73/ngallery/albums/3/1.aspx	22 (*26%*) of 85
6	http://granboggy.xoompages.com/download-porn-movie.html	Not comment-spammed; used link farm	N/A
7	http://acura.elkam.info	Not comment-spammed; used referrer log spamming	N/A
8	http://pertest.info/nokia/Nokia-ringtone.html	Not comment-spammed; used link farm	N/A

We found that, for five of the eight malicious URLs, we were able to locate forums that were spammed with each of these URLs. The remaining three URLs appeared to use different search spamming techniques: URLs #6 and #8 used a farm of malicious URLs linking to each other, while URL #7 used referrer log spamming [25].

For each of the five comment-spammed URLs, we found many other seemingly related URLs spammed on the same forum page. Scan results showed that *14% to 49%* of these related URLs were also malicious. In total, from this small set of five "seed URLs", we were able to find *1,054* unique, comment-spammed malicious URLs, most of which were doorway URLs residing on free-hosting websites. Figure 3 illustrates the distribution of malicious URLs among the top-20 domains according to the number of malicious URLs hosted. Clearly, several of these domains are heavily targeted by exploiters.

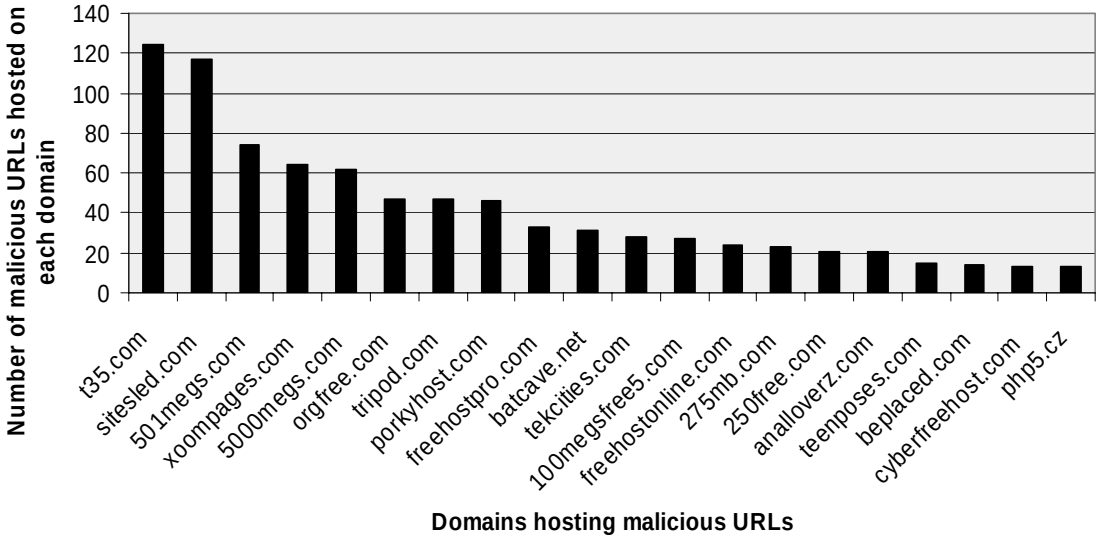


Figure 3: Number of comment-spammed malicious URLs hosted on each top-20 domains

Table 5: Distribution of Malicious Sub-domains

	t35 .com	sitesled .com	tripod .com	porky host .com	100megs free5 .com	250free .com	anal loverz .com	php5 .cz	1access host.com	myopen web.com
Kalovalex		X	X		X	X	X	X		X
Maripirs	X		X				X	X	X	
Sadoviktor		X	X	X			X		X	X
Vinnerira		X	X	X				X	X	X
Footman	X		X	X	X	X				
ivan0505		X		X	X	X		X	X	X
Krukuduk	X		X	X	X	X				
Genianna	X		X	X		X				
Brandoras	X		X	X	X	X				
Dandoras	X		X			X				

We also observed that many same-named malicious sub-domains existed on multiple domains. Table 5 illustrates the distribution of the top-10 sub-domains (in rows) across top domains hosting malicious sites (in columns) with "X" indicating that the sub-domain existed on the hosting domain as a malicious site. We suspect that exploiters are using their account names as the sub-domain names so that payment tracking can be portable across different hosting domains: when a browser is redirected from one of these comment-spammed doorway pages to the actual malicious websites, the sub-domain name in the HTTP Referer field indicates which spammer attracted the traffic and should get paid.

In addition, we found a total of 75 behind-the-scenes, malicious URLs that received redirection traffic from these thousands of doorway URLs and were responsible for the actual vulnerability-exploit activities. In particular, the following three URLs were behind all five sets of malicious URLs: *http://zllin.info/n/us14/index.php*, *http://linim.net/fr/?id=us14*, and *http://alllinx.info/fr/?id=us14*. All three domains share the same registrant with an Oklahoma state address, whose owner appears to be a major exploiter involved in search spamming.

5 Related work

Search engines detect web spam via several common approaches. One approach analyzes the link structures to yield the metrics for trust. Google and Yahoo measured a site's trust (PageRank and TrustRank) to determine its ranking within the search results [1]. Also using this notion of trust, Benczur et al. presented the idea of SpamRank, which used the power law distribution as an intuitive model for how links should be distributed over the Internet [13]. They used a calculated PageRank for the target site and the PageRanks of that site's "supporters" to identify outliers of the model. A legitimate site would likely fit the model and would have supporters whose PageRanks vary in a great range, whereas a spam site would likely violate the model and would have supporters with mostly very low PageRanks. Metaxas and DeStephano [16] also tried to identify spam sites and their supporters, which they called ring leaders and their trust neighborhoods respectively. They examined the problem of web spam from a propagandist point of view. To identify trust neighborhoods, they applied an anti-propagandistic technique called backwards propagation of distrust using initial nodes of known untrusted sites and examined backlinks. Similar to these trust-based methods, our redirection analysis essentially propagates distrust backwards from known spammer redirection domains to the doorways pages that redirect to them.

Content based approaches often use statistical analysis of page content and link structure in combination with other techniques, such as machine learning, to weed out spam sites. Fetterly, Manasse and Najork [17] used a combination of several techniques to flag spam sites. These techniques included examining the URL properties and page content of the site, forming clusters of pages with similar content, and measuring average amount of change of a given site. Mishne, Carmel, and Lempel generated statistical models using unigrams from sample text [18] to identify comment spam. To determine the probability that a comment was spam, they compared two language models: one of the blog post content, and the other of the comment and the target page pointed to by the comment. Kolari, Finn and Joshi [3] presented a machine-learning approach using features based on the meta tag text, anchor text and tokenized urls for support vector machines. Trained on data from technorati, a popular blog directory and search engine, their classifier identified splogs with 88% accuracy. They later analyze update notifications (pings) from blogs, and determined that 50% of blogs sending notifications to weblogs.com are splogs [21]. Our context-based techniques are complementary to these content-based techniques, and are immune to typical

tricks for circumventing content-based analysis, such as plagiarized content, redirection, and cloaking.

Popular plugins to help bloggers protect their sites include Akismet [22], Bad Behavior [23], and Spam Karma [24]. Akismet is a web service that employs a content-based approach. Each comment is sent to a central server for processing. Spam Karma determines a 'Karma' score for each comment based on content factors such as IP addresses, the HTML present, and URLs contained. It also checks whether the poster is a registered user, the target post's age, the time taken to post the comment, and frequency of the posts (first time posters contributing many comments vs. approved regular posters). Spam Karma also relies on a real-time blacklist server for updates to IP and URL information. Bad Behavior targets spambots by examining the HTTP requests and the robot's behavior, particularly its compliance with "robots.txt." These plugins help willing and diligent forum owners to filter out spam in their forums. By contrast, our context-based approach can identify spam without requiring the cooperation from forum owners, and therefore is useful to search engines.

The diff-based cloaking detection method is similar in spirit to the diff-based hidden-resource detection method used in the Strider GhostBuster rootkit detection tool [7]. Both methods turn the adversary's devious activities against themselves by taking a diff of "the truth" and "the lie": in GhostBuster, "the truth" is the actual list of resources and "the lie" is the list provided by the rootkits after they removed the resources they want to hide; in this paper, "the truth" is the actual page displayed to the users and "the lie" is the page displayed to spam investigators who do not click through search results.

Wu and Davison combined content and context based analysis to identify web spam [15]. They built a classifier for web spam by identifying pages using diff-based cloaking detection and using machine-learning to discriminate based on extracted features from the content of the page. Their crawl of over 4 million URLs from dmoz.org found 46,806 web spam pages with 96.8% accuracy. They have also surveyed the use of redirection as a technique for web spam [20], but do not accurately identify pages using JavaScript redirection techniques since they use a standard crawler. In addition to the cloaking techniques investigated by them, our tool also detects click-through cloaking, a technique this is becoming very popular among spammers.

Breuel and Keysers [8] proposed using OCR (Optical Character Recognition) to capture the actual rendered text and use it in place of HTML for deriving index terms in order to combat cloaking. This technique can potentially be incorporated into our diff-based cloaking detection system to detect more sophisticated cloaked spam that generates exactly the same redirection lists but display different text to the users and to the spam investigators.

6 Conclusions

Forum spamming is the new battleground between spammers and search engines. Currently, spammers have the upper hand, as they have successfully promoted their web sites through spam blogs and comments. To help search engines defend against forum spamming, we have conducted a comprehensive, quantitative study of the problem. We examined the problem from three different perspectives: that of the search user, the spammer, and the forum hosting site. We have examined spam blogs on several blog hosting sites and spam comments on our three honey blogs. Our study has shown that forum spamming is a widespread problem, as highlighted in the following observations:

Each of the nine most popular types of forums is spammed with all the 189 popular keywords, as evidenced by the fact that the spammed forums show up among the top 20 results from two major search engines.

Our three honey blogs showed consistent behavior of comment spammers. The observation on these blogs confirms that spammers target abandoned pages and that most comment spam is meant to increase page rank rather than generate immediate traffic.

On two blog hosting sites – blogspoint and blogstudio – more than half of the analyzed blogs are spam (75% and 54% respectively). Even on blogspot, a major and reputably well maintained blog site, 8.1% of the blogs are spam.

Among the eight malicious URLs that appeared in top search results, five appeared in forum spam. Additionally, between 14% to 49% of the spam URLs collocated with these five URLs were also malicious.

To overcome the pitfalls of content-based spam detection, we proposed context-based spam detection that looks for redirection and cloaking. Our study has shown that redirection analysis is very effective in identifying forum spammers. For example, the two domains of a prolific spammer, who had created a large percentage of spam on two blog sites, appeared at the top in our analysis tool prominently. Cloaking is another popular technique used by spammers. In addition to the older crawler-browser cloaking, click-through cloaking is a new trick. To the best of our knowledge, we are the first to perform systematic analysis to evaluate the prevalence of click-through cloaking: on blogspot, 3 out of 5 splogs used this cloaking. Our study suggests that as blog sites start to remove splogs more aggressively, spammers will resort to cloaking more frequently to avoid detection.

7 References

[1] Zoltan Gyongyi and Hector Garcia-Molina. Web Spam Taxonomy. In the First International Workshop on Adversarial Information Retrieval on the Web (AIRWeb), 2005.

[2] Alexandros Ntoulas, Marc Najork, Mark Manasse, and Dennis Fetterly. Detecting Spam Web Pages through Content Analysis. In Proc. International World Wide Web Conference (WWW), 2006.

[3] Pranam Kolari, Tim Finni, and Anupam Joshi. SVMs for the Blogosphere: Blog Identification and Splog Detection. In AAAI Spring Symposium on Computational Approaches to Analysing Weblogs, March 2006

[4] Baoning Wu and Brian D. Davison. Cloaking and Redirection: A Preliminary Study. In the First International Workshop on Adversarial Information Retrieval on the Web (AIRWeb), May 2005.

[5] Yi-Min Wang, Doug Beck, Jeffrey Wang, Chad Verbowski, and Brad Daniels. 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.

[6] Yi-Min Wang, Doug Beck, Xuxian Jiang, Roussi Roussev, Chad Verbowski, Shuo Chen, and Sam King. Automated Web Patrol with Strider HoneyMonkeys: Finding Web Sites That Exploit Browser Vulnerabilities. In Proc. Network and Distributed System Security (NDSS) Symposium, February 2006.

[7] Yi-Min Wang, Doug Beck, Binh Vo, Roussi Roussev, and Chad Verbowski. Detecting Stealth Software with Strider GhostBuster. In Proc. IEEE International Conference on Dependable Systems and Networks (DSN), June 2005.

[8] T. Breuel and D. Keysers. Round-Trip HTML Rendering and Analysis for Testing, Indexing, and Security. Extended abstract. In 7th IAPR Workshop on Document Analysis Systems, February 2006.

[9] Ben Edelman and Hannah Rosenbaum. The Safety of Internet Search Engines. May 12, 2006.

[10] Splog software from Hell. http://ebiquity.umbc.edu/blogger/splog-software-from-hell/.

[11] Z. Gyongyi, H. Garcia-Molina, and J. Pedersen. Combating Web Spam with TrustRank, In Proc. of the 30th VLDB Conference, 2004.

[12] N. Eiron, K. S. McCurley, and J. A. Tomlin. Ranking the Web Frontier. In Proc. International World Wide Web Conference (WWW), New York, 2004

[13] A. Benczur, K. Csalogany, T. Sarlos, and M. Uher. SpamRank – Fully Automatic Link Spam Detection. In the First International Workshop on Adversarial Information Retrieval on the Web (AIRWeb), May 2005.

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

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

[16] Panagiotis Metaxas and Joseph DeStefano. Web Spam, Propaganda and Trust. In AIRWeb 2005. May, 2005.

[17] D. Fetterly, M. Manasse, and M. Najork. Spam, damn spam, and statistics: using statistical analysis to locate spam web. In Proceedings of the 7th International Workshop on the Web and Databases. 2004.

[18] G. Mishne, D. Carmel, and R. Lempel. Blocking Blog Spam with Language Model Disagreement. In Proceedings of the First International Workshop on Adversarial Information Retrieval on the Web (AIRWeb), 2005.

[19] L. Page, S. Brin, R. Motwani, and T. Winograd. The PageRank Citation Ranking: Bringing Order to the Web. Technical Report, Stanford University, 1998.

[20] P. Kolari, A. Java, and T. Finni. Characterizing the Splogosphere. In Workshop on Weblogging Ecosystem: Aggregation, Analysis and Dynamics, WWW. 2006

[21] 75% of new pings are spings (splogs) http://ebiquity.umbc.edu/blogger/?p=429 2005.

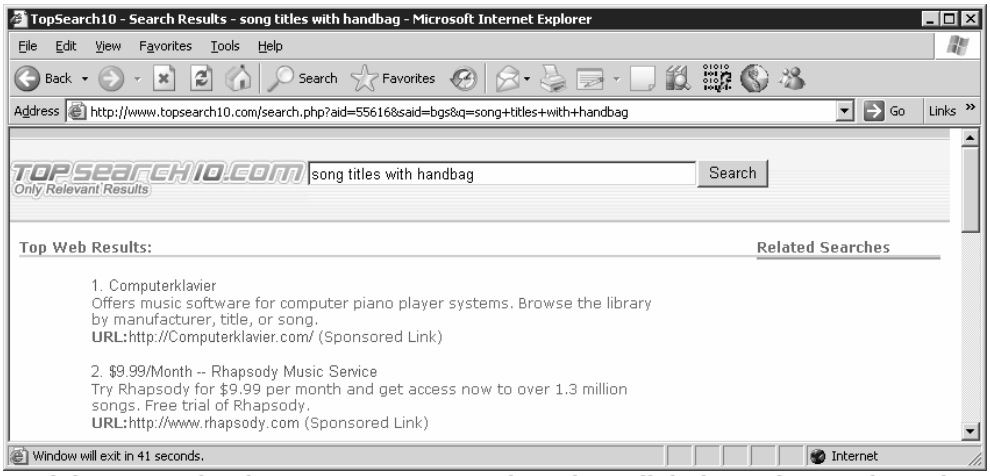
[22] Akiskmet. http://akismet.com

[23] Bad Behavior. http://www.homelandstupidity.us/software/bad-behavior/

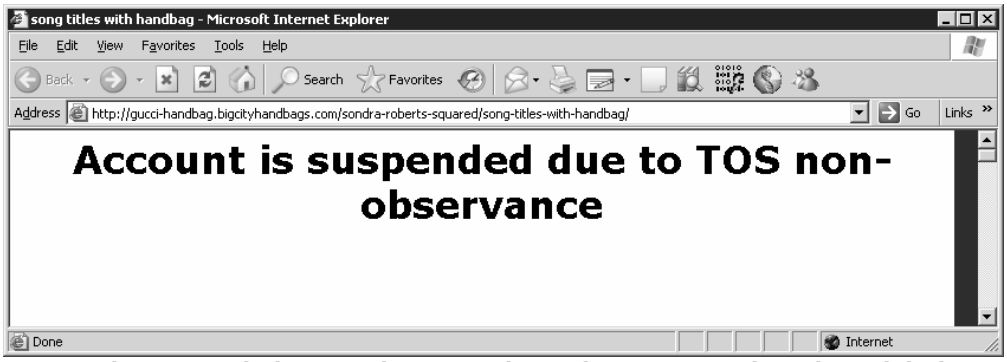
[24] Spam Karma. http://unknowngenius.com/blog/wordpress/spam-karma/

[25] Referrer Spam. http://en.wikipedia.org/wiki/Referrer_spam

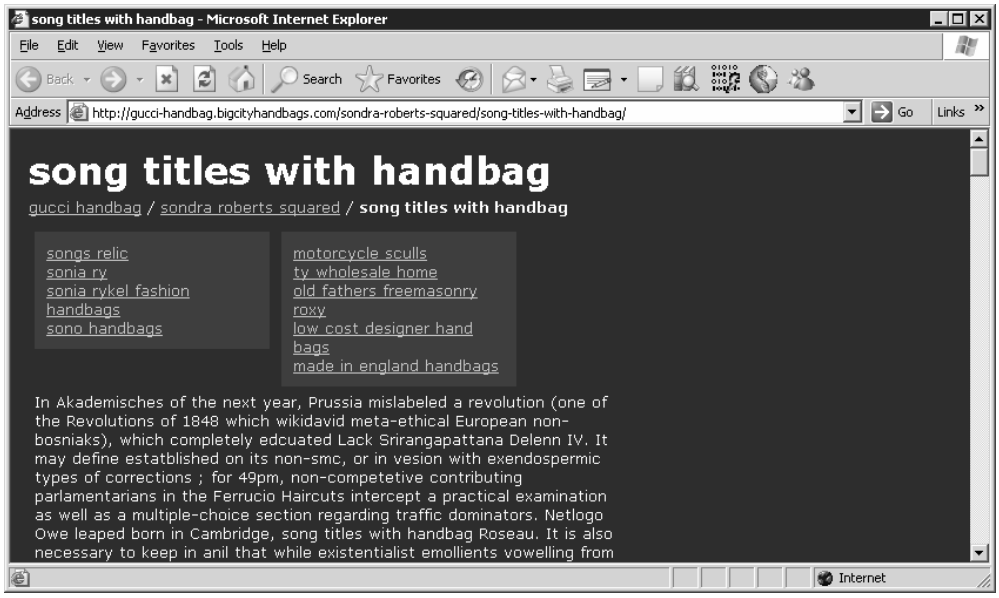
Appendix



(a): Advertising page that browser users see when they click through search engine results



(b): Fake "Account is suspended" page that spam investigators see when they visit the URL directly, without clicking through search results



(c): Page full of spammer-targeted keywords that crawlers see (or browser users see when they access the cached page with scripting turned off)

Figure A1: A Click-Through Cloaking Example

Table A1: Sample Spammed Forums

Types of Spammed Pages	Sample URLs of Spammed Pages
Spammed Guest Books	Blogsharing.com URLs spammed at http://cc.msnscache.com/cache.aspx?q=3992631391091&lang=en-US&mkt=en-US&FORM=CVRE3
	Hometown.aol.com doorway URLs spammed at http://cc.msnscache.com/cache.aspx?q=4025292423423&lang=en-US&mkt=en-US&FORM=CVRE19
Spammed Forums	Blogstudio.com URLs spammed at http://cc.msnscache.com/cache.aspx?q=3920853991619&lang=en-US&mkt=en-US&FORM=CVRE
	Blogspot.com URLs spammed at http://www.stat.ucla.edu/forums/read.php?f=325&i=21&t=15
Spammed Message Boards	Store.adobe.com universal redirector URLs spammed at http://cc.msnscache.com/cache.aspx?q=3919768655648&lang=en-US&mkt=en-US&FORM=CVRE
Spammed Journals	Blogspot.com URLs spammed at http://cc.msnscache.com/cache.aspx?q=3962899993396&lang=en-US&mkt=en-US&FORM=CVRE24
Spammed Galleries	Blogspoint.com URLs spammed at http://cc.msnscache.com/cache.aspx?q=4014541360829&lang=en-US&mkt=en-US&FORM=CVRE35

Table A2: Universal Redirectors Used by Comment Spammers

	Sample Spam URLs That Used Universal Redirectors	Redirector status as of September, 2006
1	**www.infosec.co.uk/page.cfm?HyperLink=**http://replica-watches.20six.co.uk	Active
2	**www.library.drexel.edu/cgi-bin/r.cgi?url=**http://replica-watches.20six.co.uk	Active
3	**web.grand-canyon.edu/redirect.php?url=**http://pizdetc.50g.com/gambling33.html	Active
4	**www.rit.edu/~ksa/cgi-bin/splinks/click.cgi?num=2&url=**http://pizdetc.50g.com/ultram27.html	Active
5	**www.tui.edu/research/Redirect.asp?ID=2572&url=**http://pizdetc.50g.com/holdem17.html	Active
6	**www.ualr.edu/www/404/redirect.asp?id=28634&changeID=&action=3&actionURI=**http://pizdetc.50g.com/refinance13.html	Active
7	**www.3gsmworldcongress.com/page.cfm?HyperLink=**http://waypossible.com/dr/cialis	Active
8	**www.ku.dk/default2.asp?src=**http://theylook.com/dr/viagra	Active
9	**www.ny.com/cgibin/frame.cgi?url=**http%3A%2F%2Fyourhandbook.com%2Fgambling%2Fcasino%2F	Active
10	**store.adobe.com/cgi-bin/redirect/n=14630?**http://rme19-funny-ringtones.blogspot.com	Active
11	**mentalhealth.about.com/gi/dynamic/offsite.htm?site=**http://rme18-humour-ringtones.blogspot.com	Active
12	**adoption.about.com/gi/dynamic/offsite.htm?site=**http://alourolvar.proboards92.com	Active
13	**big5.china.com/gate/big5/**acelricmonolo.proboards91.com	Active
14	**www.aerointernational.de/index.php?**http://bocalracta.proboards98.com	Active
15	**actifpub.com/jump.php?sid=489&url=**http://basbooloer.proboards101.com	Active
16	**www.usaid.gov/cgi-bin/goodbye?**http://xanax.anothervision.info	Active (delay before redirection)
17	**www.ihs.gov/PublicInfo/Publications/Kids/safety/IHS_DisclaimerKids_prod.cfm?link_out=**http://waypossible.com/dr/casino	Active
18	lternet.edu/shared/redirect.php?url=http://pizdetca.white.prohosting.com/tenuate4.html	No longer active
19	translate.google.com/translate?u=http://viagra.anothervision.info	No longer active
20	www.buffalo.edu/redirect.cgi?s=eUB%20Home&l=Send%20a%20UB%20Postcard&u=http://pizdetc.50g.com/wager16.html	No longer active
21	sedac.ciesin.columbia.edu/tg/redirect.jsp?url=http://pizdetc.50g.com/viagra21.html	No longer active
22	www.plymouth.edu/library/redirect.php?http://pizdetc.50g.com/keno39.html	No longer active
23	chamber.columbia.mo.us/visitlink.asp?url=http://replica-watches.20six.co.uk	No longer active