# Juice: A Longitudinal Study of an SEO Botnet

David Y. Wang, Stefan Savage, and Geoffrey M. Voelker
University of California, San Diego

## Abstract

*Black hat search engine optimization (SEO) campaigns attract and monetize traffic using abusive schemes. Using a combination of Web site compromise, keyword stuffing and cloaking, a SEO botnet operator can manipulate search engine rankings for key search terms, ultimately directing users to sites promoting some kind of scam (e.g., fake anti-virus). In this paper, we infiltrate an influential SEO botnet, GR, characterize its dynamics and effectiveness and identify the key scams driving its innovation. Surprisingly, we find that, unlike e-mail spam botnets, this botnet is both modest in size and has low churn—suggesting little adversarial pressure from defenders. Belying its small size, however, the GR botnet is able to successfully "juice" the rankings of trending search terms and, during its peak, appears to have been the dominant source of trending search term poisoning for Google. Finally, we document the range of scams it promoted and the role played by fake anti-virus programs in driving innovation.*

## 1 Introduction

Traffic is the lifeblood of online commerce: eyeballs equal money in the crass parlance of today's marketers. While there is a broad array of vectors for attracting user visits, Web search is perhaps the most popular of these and is responsible for between 10 and 15 billion dollars in annual advertising revenue [1, 2].

However, in addition to the traffic garnered by such *sponsored* search advertising, even more is driven by so-called "organic" search results. Moreover, it is widely held that the more highly ranked pages—those appearing at the beginning of search results—attract disproportionately greater volumes of visitors (and hence potential revenue). Thus, a large ecosystem has emerged to support *search engine optimization* or SEO—the practice of influencing a site's ranking when searching under specific query terms. Many of these practices are explicitly encouraged by search engines with the goal of improving the overall search experience (e.g., shorter load times, descriptive titles and metadata, effective use of CSS to separate content from presentation, etc.) and such approaches are commonly called "white hat" SEO techniques. However, on the other side of the spectrum are "black hat" techniques that explicitly seek to manipulate the search engine's algorithms with little interest in improving some objective notion of search quality (e.g., link farms, keyword stuffing, cloaking and so on).

Unsurprisingly, such black hat techniques have quickly been pressed into the service of *abusive advertising*—advertising focused on attracting traffic for compromise (e.g., drive-by downloads [4]), for fraud (e.g., fake anti-virus [17]), or for selling counterfeit goods (e.g., pharmaceuticals or software).[1] While a few such incidents would not generate alarm, there is increasingly clear evidence of large-scale SEO campaigns being carried out: large numbers of compromised Web sites harnessed in unison to poison search results for attractive search queries (e.g., trending search terms). Indeed, one recent industry report claims that 40% of all malware infestations originate in poisoned search results [9]. However, the details of how such *search poisoning attacks* are mounted, their efficacy, their dynamics over time and their ability to manage search engine countermeasures are still somewhat opaque.

In service to these questions, this paper examines *in depth* the behavior of one influential search poisoning botnet, "GR".[2] In particular, we believe our work offers three primary contributions in this vein.

*Botnet characterization.* By obtaining and reverse engineering a copy of the "SEO kit" malware installed on compromised Web sites, we were able to identify other botnet members and infiltrate the command and control channel. Using this approach we characterize the activities of this botnet and its compromised hosts for nine months. We show that unlike email spamming botnets, this search poisoning botnet is modest in size (under a thousand compromised

[1] Indeed, in one recent study of counterfeit online pharmaceuticals the most successful advertiser was not an email spammer, but rather was an SEO specialist [14].

[2] Each of the functions and global variables in this botnet are prefixes with a capital GR. We believe it is an acronym, but at the time of this writing we do not know what the authors intended it to stand for.

Web sites) and has a low rate of churn (with individual sites remaining in the botnet for months). Moreover, we document how the botnet code is updated over time to reflect new market opportunities.

*Poisoning dynamics.* By correlating captured information about the keywords being promoted with contemporaneous Internet searches, we are able to establish the effectiveness of such search poisoning campaigns. Surprisingly, we find that even this modest sized botnet is able to effectively "juice" the ranking of thousands of specific search terms within 24 hours and, in fact, it appears to have been the dominant contributor to poisoned trending search results at Google during its peak between April and June 2011.

*Targeting.* By systematically following and visiting the "doorway" pages being promoted, both through redirections and under a variety of advertised browser environments, we are able to determine the ultimate scams being used to monetize the poisoning activity. We find evidence of a "killer scam" for search poisoning and document high levels of activity while the fake antivirus ecosystem is stable (presumably due to the unusually high revenue generation of such scams [17]). However, after this market experienced a large setback, the botnet operator explores a range of lower-revenue alternatives (e.g., pay-per-click, drive-by downloads) but never with the same level of activity.

Finally, in addition to these empirical contributions, our paper also documents a methodology and measurement approach for performing such studies in the future. Unlike email spam which delivers its content on a broad basis, search poisoning involves many more moving parts including the choice of search terms and the behavior of the search engine itself. Indeed, our analyses required data from three different crawlers to gather the necessary information: (1) a host crawler for identifying and monitoring compromised Web sites, (2) a search crawler to identify poisoned search results and hence measure the effectiveness of the poisoning, and (3) a redirection crawler that follows redirection chains from doorway pages linked from poisoned search results to identify the final landing pages being advertised.

The remainder of this paper is structured as follows. In Section 2, we walk through an example of a search poisoning attack and explain how our study builds on prior work. In Section 3 we describe the GR SEO botnet in detail, followed by a description of Odwalla, the system we built to monitor and probe its activities in Section 4. Finally, we describe our analyses and findings in Section 5, summarizing the most cogent of these in our conclusion.

# 2 Background

As background, we start with an example of a search poisoning attack and then discuss previous work that has explored the effects of search engine poisoning.

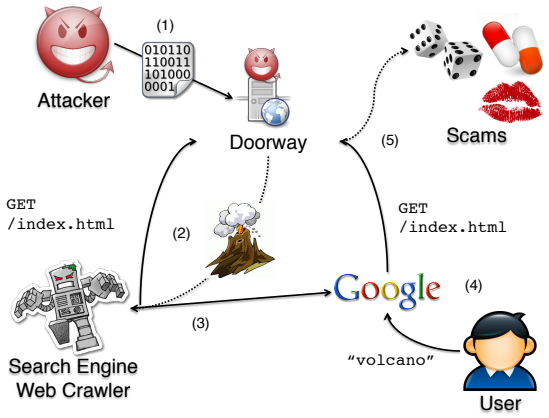



**Figure 1: A typical search poisoning attack.**

## 2.1 An Example

Figure 1 shows the steps of a typical search poisoning attack, which baits users into clicking through a search result to be redirected to a scam. In this example, we presuppose that due to exogenous factors there is sudden interest in terms related to volcanoes (e.g., an eruption somewhere). The scam proceeds as follows: (1) The attacker exploits a vulnerability on a Web site and installs an SEO kit (Section 3), malware that runs on the compromised site and changes it from a legitimate site into a *doorway* under the attacker's control. (2) Next, when a search engine Web crawler requests the page `http://doorway/index.html`, the SEO kit detects the visitor as a crawler and returns a page related to volcanoes (the area of trending interest) together with cross links to other compromised sites under the attacker's control. (3) The search engine indexes this page, and captures its heavy concentration of volcano terms and its linkage with other volcano-related sites. (4) Later a user searches for "volcano" and clicks through a now highly ranked search result that links to `http://doorway/index.html`. (5) Upon receiving this request, the SEO kit detects that it is from a user arriving via a search engine, and attempts to monetize the click by redirecting the user to a scam such as fake AV.

## 2.2 Previous Work

Previous work, dating back well over a decade, has studied cloaking mechanisms and Web spam in detail [12, 19, 20, 21]. Recently, interest has focused on measuring the phenomenon of search result poisoning and the resulting negative user experience, together with various methods for detecting poisoned search results as a step towards undermining the attack. In this paper we extend this line of work by characterizing the coordinated infrastructure and organization behind these attacks from the attacker's point of view,

and the strategies an attacker takes both in monetizing user traffic as well as responding to intervention.

For example, Wang et al. recently measured the prevalence of cloaking as seen organically by users in Web search results over time for trending and pharmaceutical queries [19]. Cloaking is a "bait and switch" technique where malware delivers different semantic content to different user segments, such as SEO content to search engines and scams to users, and is one of the essential ingredients for operating a modern black hat SEO campaign. Similarly, Lu et al. developed a machine learning approach for identifying poisoned search results, proposing important features for statistical modeling and showing their effectiveness on search results to trending terms [12]. During the same time period, Leontiadis et al. [10] and Moore et al. [15] also measured the exposure of poisoned search results to users, and used their measurements to construct an economic model for the financial profitability of this kind of attack. Despite the common interest in search result poisoning, these studies focus on how cloaking was utilized to manipulate search results and its impact on users, whereas our work focuses more on the mechanisms used by and the impact of an entire SEO campaign coordinated by an attacker via a botnet.

The work of John et al. is the most similar to the study we have undertaken [5]. Also using an SEO malware kit, they extrapolated key design heuristics for a system, *deSEO*, to identify SEO campaigns using a search engine provider's Web graph. They found that analyzing the historical links between Web sites is important to detecting, and ultimately preventing, SEO campaigns. Our work differs in that, while we study a similar SEO kit, we focus on the longitudinal operation of SEO campaigns as organized by an SEO botnet operator: what bottlenecks, or lack thereof, an operator faces, and what factors, such as interventions, appear to have influenced the operator's behavior over time.

# 3 The GR Botnet

In this section we present the architecture of the GR botnet responsible for poisoning search results and funneling users, as traffic, to various scams. We start by introducing its SEO malware kit, and then present a high-level overview of its architecture, highlighting specific functionality found in the SEO kit and the evolution of the source code.

## 3.1 SEO Kit

An SEO kit is software that runs on each compromised Web site that gives the botmaster backdoor access to the site and implements the mechanisms for black hat search engine optimization. We obtained an SEO kit after contacting numerous owners of compromised sites. After roughly 40 separate attempts, one site owner was willing and able to send us the injected code found on their site. Although we cannot pinpoint the original exploit vector on the compromised Web site, there have been many recent reports of attackers compromising Web sites by exploiting Wordpress and other similar open source content management systems [13].

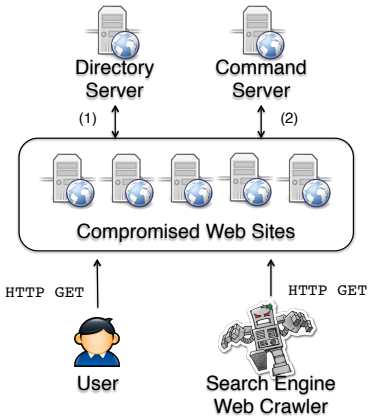



**Figure 2: A user and a search engine Web crawler issue a request to a compromised Web site in the botnet. The site will (1) contact the directory server for the address of the C&C, and then (2) contact the C&C for either the URL for redirecting the user, or the SEO content for the Web crawler.**

The SEO kit is implemented in PHP and consists of two components, the *loader* and the *driver*. The loader is initially installed by prepending PHP files with an `eval` statement that decrypts base64 encoded code. When the first visitor requests the modified page, causing execution of the PHP file, the loader sets up a cache on the site's local disk. This cache reduces network requests, which could lead to detection or exceeding the Web site host's bandwidth limits. Then the loader will contact a directory server using an HTTP GET request to find the location of a command-and-control server (C&C) as either a domain name or IP address. Upon contacting the C&C server, the loader downloads the driver code which provides the main mechanisms used for performing black hat SEO.

## 3.2 Botnet Architecture

Figure 2 shows the high-level architecture of the botnet. The botnet has a command and control architecture built from pull mechanisms and three kinds of hosts: compromised Web sites, a directory server, and a command and control server (C&C).

### 3.2.1 Compromised Web Sites

Compromised Web sites act as doorways for visitors and are controlled via the SEO kit installed on the site. The SEO

kit uses *cloaking* to mislead search engines, users, and site owners, as well as to provide a control mechanism for the botmaster. Cloaking is a mechanism that returns different content to different types of users based upon information gleaned from the HTTP request (Figure 2).

The SEO kit first checks to see if the user is a search engine crawler. If it is, the SEO kit returns content to the crawler to perform black hat search engine optimization. When the SEO kit is invoked via an HTTP GET request, the driver looks up the hostname of the visitor's IP address using `gethostbyaddr`. It then searches for the substring `googlebot.com` within the hostname to determine if Google's search crawler is accessing the page.[3] If the match is successful, the driver pulls SEO content from the C&C server and returns it to the crawler with the specific goal of improving the ranking of the page in search results independent of the original content of the page. Specifically, the driver builds a page with text and images related to the trending search results that link to the site. The SEO kit retrieves this content on demand by issuing auxiliary requests to search engines and *spinning* content constructed from the resulting search results snippets and images.[4]Additionally, the SEO kit inserts links to other nodes of the botnet, as directed by the C&C, into the spun content to manipulate the search engine's ranking algorithms. As search engines typically use the number of backlinks to a page as one signal of high desirability [18], the botmaster aims to develop a linking strategy to improve the ranking of compromised sites in the SEO botnet.

If the SEO kit does not identify the visitor as a crawler, the driver next checks if the visit reflects user search traffic. The SEO kit identifies users by reading the `Referrer` field in the HTTP request headers, and verifying that the user clicked through a Google search results page before making the request to the compromised site. For these users, the SEO kit contacts the C&C server on demand for a target URL that will lead users to various scams, such as fake anti-virus, malware, etc., all of which can earn money for the botmaster. The SEO kit then returns this target URL together with redirect JavaScript code as the HTTP response to trigger the user's browser to automatically visit the target.

The SEO kit also uses its cloaking mechanism to provide backdoor access to the compromised site for the botmaster. To identify the botmaster, the SEO kit inspects the `User-Agent` field in the HTTP request headers, looking for a specific, unique phrase as the sole means of authentication. With this authentication token, the botmaster has the ability to read files from the local hard disk of the site, fetch URLs while using the compromised site as a proxy, run scripts pulled from the C&C, etc., all controlled through parameters to HTTP GET requests.

Finally, if the visitor does not match either the Googlebot crawler, a user clicking on a search result, or the backdoor, then the SEO kit returns the original page from the site before it was compromised. Thus, site owners who visit their pages directly will be unaware of the compromise.

[3]It appears that the botmaster is only interested in poisoning Google's search results, as they solely target the Googlebot crawler—a trend also observed in previous cloaking studies [19].

[4]Spinning is another black hat SEO technique that rephrases and rearranges text to avoid duplicate content detection.

#### 3.2.2 Directory Server

The directory server's only role is to return the location of the C&C server, either as a domain or IP address. Although relatively simple in functionality, it is the first point of contact from the compromised Web sites in the botnet and performs the important function of rendezvousing a compromised site with the C&C server. As a result, the directory server must be reachable and available and the SEO kit uses a typical multi-step process to locate it. The SEO kit will first attempt to reach the directory server through a hard-coded domain from the SEO kit, then a hard-coded IP address, before finally resorting to a backup domain generation algorithm (DGA) calculated using a time-based function. The directory server appears to have received little takedown pressure, though. We probed the potential backup domains up to a year into the future and found that no backup domains were registered, suggesting that this final fallback has not been necessary.

#### 3.2.3 Command Server

The C&C server acts as a centralized content server where the botmaster stores data that the compromised sites will eventually pull down. The content is mostly transient in nature, and includes the trending search terms to target with SEO, the redirect URLs returned to users leading them to scams, and even the driver component of the SEO kit. This architecture allows the botmaster to make a single update that eventually propagates to all active nodes of the botnet.

### 3.3 SEO Kit Evolution

Examining the SEO kit's source revealed a variety of comments in the code. These comments were primarily written in Russian, suggesting the SEO campaign is implemented and operated by Russian speakers. From the translated comments we saw hints of the existence of revious versions of the SEO kit in the wild, such as:

```
/**
 * v7.2 (14.09.11)
 * - Automatic cleaning of other malware
 *
 * v7.1 (05.09.11)
 * - Re-written for object oriented model
```

| Date | Version | Capability |
|---|---|---|
| Aug 6 2010 | page v1 | Build SEO page using Bing search results.<br>User-Agent cloaking against Google, Yahoo, and Bing while ignoring "site:" queries.<br>Redirect traffic from Google, Yahoo, Bing search using JS through `gogojs.net`. |
| Sep 22 2010 | index v1.1 | Reverse DNS cloaking against Googlebot. |
| Oct 6 2010 | page v2.1 | Use statistical model (# links, # images) to build SEO page.<br>Also redirect traffic from Google Image Search.<br>Redirect traffic with HTTP 30X and use cookie to redirect only once a day per visitor. |
| Mar 29 2011 | page v4 | Modify .htaccess to rewrite URLs and use Google Suggest terms for cross linking.<br>Reverse DNS cloaking only against Googlebot. |
| Jul 15 2011 | index v6 | Hotlink images from Bing Image Search to help build SEO page. |
| | page v5 | Proxy images instead of hotlinking. |
| Aug 18 2011 | v7 | index + page code branches merged.<br>Morph proxied images.<br>Redirect traffic using JS. |
| Sep 14 2011 | v7.2 | Clean other malware. |
| Sep 27 2011 | vOEM | OEM terms targeted. |
| Oct 28 2011 | vMAC | Mac OS X OEM terms targeted for low frequency traffic.<br>Redirect traffic from any Google service due to referer policy change. |
| Mar 06 2012 | v8 | Only redirect Google Image Search traffic. |

**Table 1: Timeline of SEO kit versions along with the capabilities added in each version. The SEO techniques used are colored blue. The redirect mechanisms and policies for funneling traffic are colored purple. The various cloaking techniques and policies are colored green. And orange capabilities focus specifically on Google Image Search poisoning. The remaining are purely informational.**

These indications of previous versions of the SEO kit motivated us to search for them using identifying substrings unique to the SEO kit code, such as "GR_HOST_ID". We discovered that previous versions were posted on the Web by site owners who were seeking assistance in deciphering the injected code on their site. After verifying older versions existed, we were able to download additional previous versions of the SEO kit from the C&C server by reverse engineering the protocol for downloading the driver and fuzzing likely inputs. In the end, we were able to download nearly all major SEO kit revisions since August 2010.

As seen in the sample above, the comments from each version of the SEO kit have a date and a short log message about the update similar to a version control system. From these comments, we reconstructed the developments in the SEO kit and thus the evolution of the SEO botnet and the botmaster's SEO strategies over two years. Table 1 summarizes our findings by presenting changes in capabilities with the corresponding version and date. Below are some highlights, many of which confirmed our early theories.

**Structure.** The compromised sites were at one time divided into *indexers*, which SEO-ed search engine visitors, and *doorways*, which redirected users, each with different cloaking mechanisms and policies. Starting August 2011, however, the code was merged into a single SEO kit with a unified cloaking mechanism and policy.

**Cloaking.** Initially, the doorways and indexers used User-Agent cloaking, where the server examines the `User-Agent` field in the HTTP request headers to identify user traffic and avoid detection. Specifically, the doorways used the cloaking mechanism to identify visitors who clicked through one of the three largest search engines: Google, Yahoo, Bing. By late September 2010, however, the indexers implemented the reverse DNS cloaking mechanism as described above. Similarly, by late March 2011 the doorways used the same cloaking mechanism and began targeting user traffic from Google exclusively.

**Redirection.** The redirection mechanism, used to funnel user traffic to scams, also changes significantly over time. Originally, the doorways redirected user traffic using JavaScript through an intermediary site, `gogojs.net`, which we suspect served as a traffic aggregation hop to collect statistics. By October 2010, the doorway redirected traffic via the HTTP 30* status with a cookie to limit visitors to one visit per day. Then in August 2011, the SEO kit returns to using JavaScript to redirect user traffic.

**SEO.** The SEO models and policies, used by the SEO kit to manipulate search result ranking, also change heavily over time. In the earliest version we have, the SEO page returned to search engine crawlers was generated from Bing search results. Then the SEO kit began using a statistical model when building an SEO page, requiring that the SEO page contents be composed of various percentages of text, images, and links. In late March 2011, the SEO kit used Google Suggest to target long-tail search terms. Then in late September 2011 it began to poison search results for OEM

queries. And by late October 2011, the SEO kit started poisoning Mac OEM queries, also long-tail search terms.

**Image Search.** One of the surprising findings from the SEO kit code is the amount of effort placed in poisoning Google Image Search. The doorways first started redirecting user traffic from Google Image Search in October 2010. In July 2011, the indexers hotlinked images from Bing to help build the SEO page and shortly thereafter the doorways began proxying images instead of hotlinking. By August 2011, the SEO kit began morphing the images, such as inverting them, to avoid duplicate detection. And currently, since March 2012, the SEO kit only redirects traffic from Google Image Search.

# 4 Methodology

We use data from three crawlers to track the SEO botnet and monitor its impact: (1) a botnet crawler for tracking compromised Web sites in the botnet and downloading SEO data from the C&C server, (2) a search crawler that identifies poisoned search results in Google, enabling us to evaluate the effectiveness of the botnet's black hat SEO, and (3) a redirection crawler that follows redirection chains from the doorway pages linked from poisoned search results to the final landing pages of the scams the botmaster uses to monetize user traffic. Table 2 summarizes these data sets, and the rest of this section describes each of these crawlers and the information that they provide.

## 4.1 Odwalla Botnet Crawler

We implemented a botnet crawler called Odwalla to track and monitor SEO botnets for this study. It consists of a host crawler that tracks compromised Web sites and a URL manager for tracking URL to site mappings.

**Host Crawler.** The host crawler tracks the compromised Web sites that form the SEO botnet. Recall from Section 3.2.1 that the SEO kit provides a backdoor on compromised sites for the botmaster through the HTTP request's `User-Agent` field. While this backdoor provides access to many possible actions, the default response is a simple diagnostic page with information about the compromised Web site such as:

```
Version: v MAC  1 (28.10.2011)
Cache ID: v7mac_cache
Host ID: example.com
```

These fields show the basic configuration of the SEO kit: the version running on the compromised site, the version of the cache it is running, and the compromised site's hostname. The diagnostic page also reports a variety of additional information, such as the relative age of the SEO kit (for caching purposes), various capabilities of the Web host (e.g., whether certain graphics libraries are installed), and information about the requestor and request URL (e.g., whether the visitor arrived via Google Search). While the majority of this information allows the botmaster to debug and manage the botnet, we use the diagnostic page to both confirm a site's membership in the botnet and monitor the status of the compromised site.

The host crawler maintains a set of potentially compromised sites together with site metadata, such as the representative probe URL for a site and the last time it confirmed the site as compromised. The probe URL is the URL that the host crawler visits for each potentially compromised site. Since a given site may have many URLs that link to different pages, all managed by the same SEO kit, the host crawler maintains one active probe URL per site to limit crawl traffic. As URLs expire, a URL manager (described below) provides alternate probe URLs for a site. The host crawler visits each probe URL twice, once to fetch the diagnostic page and once to fetch the SEO page—the page returned to search engines—containing the cross links.

The last time the site was detected as compromised influences the crawling rate. The host crawler visits all sites that were either previously confirmed as compromised, using the diagnostic page mechanism described above, or newly discovered from the cross links. It crawls these sites at a four-hour interval. For the sites that were not confirmed as compromised, for example because it could not fetch the diagnostic page, the host crawler visits them using a two-day interval as a second chance mechanism. If it does not detect a site as compromised after eight days, it removes the site from the crawling set. This policy ensures that we have near real time monitoring of known compromised sites, while limiting our crawling rate of sites where we are uncertain.

We used three methods to bootstrap the set of hosts for Odwalla to track. First, in October 2011 and then again in January 2012, we identified candidate sites using manual queries in Google for literal combinations of search terms targeted by the SEO botnet. Since the terms formed unusual combinations, such as "herman cain" and "cantaloupe", typically only SEO pages on compromised sites contained them. Second, since these pages contained cross links to other compromised sites for manipulating search ranking algorithms, we added the cross links as well. Interestingly, these cross links were insufficient for complete bootstrapping. We found multiple strongly connected components in the botnet topology, and starting at the wrong set of nodes could potentially only visit a portion of the network. Finally, we modified the SEO kit to run our own custom bots that infiltrated the botnet. These custom bots issued requests to the C&C server to download targeted search terms and links to other hosts in the botnet, providing the vast majority of initial set of bots to track. Once bootstrapped, the host

| | Odwalla | Dagger | Trajectory |
|---|---|---|---|
| **Time Range** | October 1011 – June 2012 | April 2011 – August 2011 | April 2011 – August 2011 |
| **Data Collected** | Diagnostic pages and cross links from nodes of SEO campaign. | Cloaked search results in trending searches over time. | Redirect chains from cloaked search results in trending searches. |
| **Data Perspective** | SEO Campaign botmaster. | Users of search engines. | Users of search engines. |
| **Contribution** | Characterize support infrastructure of SEO campaign. | Assess efficacy of SEO campaign. | Analyze landing scams. |

**Table 2: The three data sets we use to track the SEO botnet and monitor its impact.**

crawler used the cross links embedded in the SEO pages returned by compromised sites to identify new bots to track.

**URL Manager.** The host crawler tracks compromised sites using one probe URL to that site at a time. Often a site can have multiple pages infected with the SEO kit, though, such as a site with multiple blogs, all of the comment pages attached to blogs and articles, etc. Over time, a site owner may remove or clean an infected page while other URLs to other pages on the site remain compromised and active with the same SEO kit. In these cases, the host crawler switches to a new URL to continue to track and monitor this compromised site.

The URL manager addresses this need. It maintains a list of all URLs, as discovered from cross links, for a given site in the crawling set. It periodically checks whether each URL could potentially serve as the probe URL for a particular site by attempting to fetch a diagnostic page from that URL. Then, whenever the host crawler cannot fetch a diagnostic page for a site, it consults the URL manager to find another representative probe URL, if one exists. If not, the host crawler will continue to use the same probe URL, eventually timing out after eight days if all URLs to the site are not operational. In this case, it declares the site as "sanitized" since the SEO kit is no longer operational. Because there are far more URLs than sites, the URL manager crawls just once a day to check newly discovered URLs.

### 4.2 Dagger Search Crawler

Before we began crawling the SEO botnet, we previously explored the general dynamics of cloaking on the Web [19]. We knew from examining the code of previous versions of the SEO kit that the botnet poisoned trending search terms from April 2011 through September 2011, so we suspected that poisoned search results from the SEO botnet would also appear in our previous data set.

We had collected cloaking data using a crawler called Dagger, which ran every four hours to: (1) download trending search terms, (2) query for each trending search term on various search engines, (3) visit the page linked from each search result, and (4) run a cloaking detection algorithm to identify poisoned search results. The Dagger cloaking data allows us to analyze the impact of the SEO botnet on trending search results in near real time for a seven-month period (Section 5.3). Unfortunately, although we had continued to crawl cloaking search results, the SEO botnet changed its SEO policy to target first OEM software and then random search terms, so we would expect only accidental overlap with the Dagger data after September 2011.

### 4.3 Trajectory Redirection Crawler

While the host crawler downloads the contents of the doorway pages directly linked by poisoned search results, we also want to identify which sites these doorways ultimately lead to (e.g., a fake antivirus scam page) and hence infer how the botmaster monetizes user traffic. Since following the doorway pages to final landing pages typically involves following a complicated redirection chain, often involving JavaScript redirection code, we used the high-fidelity Trajectory crawler from yet another project [11]. This crawler uses an instrumented version of Mozilla Firefox to visit URLs, follows all application-level redirects (including JavaScript and Flash), logs the HTTP headers of the intermediate and final pages of a redirect chain, and captures the HTML and a screenshot of the final page. For all of the poisoned search results crawled by Dagger, we also crawled them using this Trajectory crawler to track scams.

## 5 Results

With the data sets we have gathered, we now characterize the activities of the SEO botnet and its compromised hosts.

### 5.1 Infrastructure

Using the nine months of data collected by Odwalla, we start by analyzing the botnet infrastructure used in the SEO campaigns: the scale of the botnet, the lifetime of compromised sites in the botnet, and the extent to which the botmaster monitors and manages the botnet.

#### 5.1.1 Scale

Compared to other kinds of well-known botnets, such as spamming botnets with tens to hundreds of thousands of

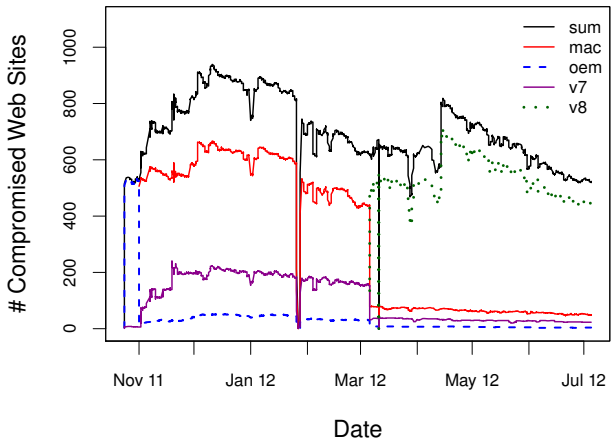



**Figure 3: Number of active nodes in the botnet over time. SUM shows the total number of active nodes, and the other lines show the number of nodes operating different versions of the SEO kit.**

hosts, the SEO botnet is only modest in size. Figure 3 presents the measured size of the botnet over time. Each line shows the number of nodes operating a specific version of the SEO kit, and the SUM line shows the total number of all nodes across all versions. For example, on December 1, 2011, we found 821 compromised sites in total, of which 585 sites were running the MAC version of the SEO kit, 42 were running OEM, and 194 were running V7.

Also unlike other kinds of botnets, the SEO botnet does not exhibit frequent churn. Over nine months, the botnet consisted of 695 active nodes on average, with a maximum size of 939 nodes on December 11, 2011. Yet, we observed the botnet running on a total of just 1,497 unique compromised sites across the entire measurement period. In contrast, spamming botnets like Storm would experience churn of thousands of hosts a day [6].

Instead, we see a few key points in time where the botnet membership fluctuates in response to SEO kit updates by the botmaster, rather than from external intervention. At the time of the upgrades, the botmaster also changes the cross linking policy among nodes, potentially revealing new nodes. In between these upgrades, the botnet size primarily fluctuates due to variations in host availability, with a degree of slow attrition. For example, on November 1, 2011, the botmaster updated the SEO kit from OEM→MAC. Even though the OEM nodes appear to have entirely switched over to MAC, the size of the botnet increases by over 200 nodes, all due to nodes running the older version V7. It appears that during the update the botmaster changed the cross linking policy to include additional nodes running V7, incidentally widening our vantage point but only for stagnant sites running an older version. March 6, 2012, marks a similar version switch over from MAC→V8 in response to another software upgrade. In this upgrade, the 298 newly discovered compromised sites were running the latest version (V8), and were discovered a month later, due to what we suspect is the deployment time for a new cross linking mechanism that utilizes `blogspot.com` as a level of indirection. Note that the large drop in botnet size on January 28, 2012, corresponds to an outage on the directory server that triggered errors on the nodes, making the nodes unresponsive to our crawler.

As a final data point, recall from Section 3.2 that the GR botnet uses a pull mechanism to ensure that compromsied sites always run an updated version of the SEO kit. As a first step, a site makes up to three attempts to contact the directory server using first a hardcoded domain, then a hardcoded IP address, and finally the output of a time-based domain generation algorithm (DGA).

While crawling the botnet we found that both the directory server's hard coded domain and IP address were unreachable starting on September 9, 2012. We took advantage of this occurrence by registering the DGA domains that compromised sites will contact when attempting to reach the directory server. Thus, we pose as the directory server and intercept all requests from the botnet's compromised sites for nearly a month between October 4 through October 30, 2012. From this vantage, we found that 1,813 unique IPs contacted our directory proxy. Since we found that, on average, 1.3 compromised sites are hosted behind a unique IP from the host crawler data, extrapolation places the botnet at 2,365 compromised sites—in agreement with our findings above that the GR botnet is on the scale of thousands of nodes.

### 5.1.2 Lifetime

The relatively stable botnet size and membership suggests that compromised sites are long-lived in the botnet. Indeed, we find that the many of these sites remain compromised for long periods of time and the botmaster is able to use them continuously without needing to constantly refresh the botnet with fresh sites to maintain viability.

We define the *lifetime* of a compromised site as the time between the first and last time the crawler observed the SEO kit running on the site. This estimate is conservative since a site may have been compromised before we first crawled it. However, we note that our measurement period of compromised sites is nine months and we began monitoring 74% of all 1497 compromised sites within the first 40 days of our study. Thus, even without the exact time of compromise, we are still able to observe the sites for long periods of time. (As further evidence, for the 537 sites that also appear in the earlier Dagger search results (Section 4.2) the majority were compromised back to April 2011.)

We decide that a site is cleaned when the site does not respond to the SEO C&C protocol for eight consecutive days,

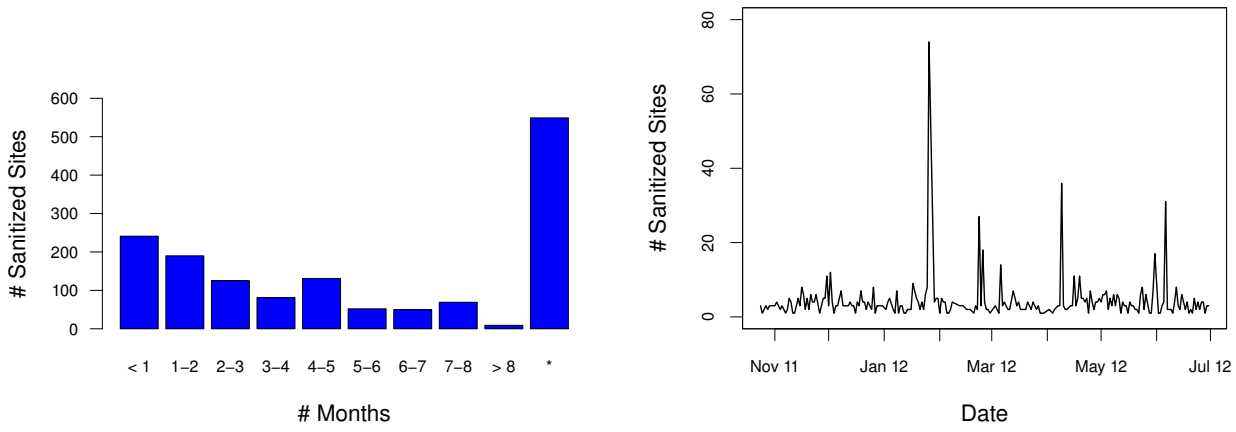



**Figure 4: On the left, the distribution of time that sites were compromised (sanitized sites only); the '*' bin shows the number of compromised sites still actively running the SEO kit at the end of the measurement period. For sites that were sanitized, the right graph shows the number of sites sanitized each day.**

suggesting that the site no longer runs the SEO kit. Typically a site stops running the SEO kit because the site owner removed the SEO malware, sanitizing the site, or the Web host or DNS registrar made the site unavailable by preventing visitors from loading the site or resolving the domain.

Consequently, the botmaster is able to use compromised sites for SEO campaigns for long periods of time. Figure 4a presents a histogram of the lifetime of the compromised sites. We distinguish between sites that have been sanitized, avoiding right-censoring of their lifetimes, and sites that have not yet been sanitized. For compromised sites that are eventually sanitized, we bin them according to their respective lifetimes using monthly intervals (30 days). Over 74% of sanitized sites have a lifetime greater than a month, and over 54% have a lifetime greater than two months. There is also a long tail, with the lifetime of some sanitized sites extending beyond even eight months. For compromised sites that have not yet been sanitized, we show them in the '*' bin. These remaining 549 sites are still compromised at the time of writing, and the majority of those have been compromised for at least seven months. This distribution indicates that the majority of compromised sites are indeed long-lived and able to support the SEO campaign for months with high availability.

Figure 4b shows the number of sites sanitized each day, indicating a low daily attrition rate of sites leaving the botnet over time (9.9 sites on average). The few spikes in the graph are specific points in time when many compromised sites were sanitized. In some cases, the spikes are partially attributable to a single entity, owning or hosting multiple sites, who cleans multiple sites at the same time. By manually comparing the resolved IP address for domain names as well as parsing WHOIS records, we were able to confirm shared hosting and shared owners, respectively. Note that the largest spike on January 26, 2012, corresponds to the outage of the botnet directory server.

One reason that sites remain compromised for long periods of time is that the SEO kit camouflages its presence to site owners. As discussed in Section 3.2.1, the SEO kit returns the original contents of the page to a visitor unless the SEO kit can determine if the visitor is a search engine crawler or has clicked on a result returned from a search engine. Hence, site owners accessing their own pages typically will not notice an installed SEO kit. That said, even when they discover the presence of the SEO kit, oftentimes they are unable or unwilling to remove it. In December and January, for instance, we contacted nearly 70 site owners to inform them that their site was infected with the SEO kit, yet just seven sites subsequently removed it.

### 5.1.3 Control

We use two different approaches to assess the botmaster's ability to monitor and manage the botnet. In the first approach, we observe what fraction of the compromised sites update their SEO kit when the botmaster deploys a new version. We can detect both version changes and site updates by parsing the version information from the diagnostic pages periodically fetched by the host crawler.

As discussed in Section 5.1.1, the data collected by the host crawler overlaps with two version updates. On November 1, 2011, version OEM updated to MAC and then, on March 6, 2012, MAC updated to V8. In both cases, we see a near instantaneous update to the respective new versions from the majority of the compromised sites, followed by a sudden addition of newly seen compromised sites.

In the OEM→MAC update, we see many *stragglers*, sites that continue running older versions of the SEO kit after the majority of sites update themselves to the latest version. Within a month after the first update, 324 out of 970 sites (33%) that comprise the botnet were stragglers. These stragglers suggest that the botmaster lacks full installation privileges on the compromised sites and is unable to force

| Group | <11/01 | 11/01 – 01/28 | 01/28 – 03/06 |
|---|---|---|---|
| <10 | 532 | 949 | 834 |
| 10 – 100 | 71 | 28 | 31 |
| 100 – 1000 | 12 | 9 | 7 |

**Table 3: The number of compromised Web sites grouped by the average amount of juice received, for the three distinct time ranges.**

an update. There is no advantage to running old versions because they poison an outdated set of search terms, are not well optimized in search results, and consequently will not attract much traffic. Therefore, the 324 stragglers represents a substantial inefficiency in the botnet. The straggler phenomenon also occurs during the second update, but the numbers are less pronounced.

Our second approach for assessing control looks at how the botmaster adjusts the cross linking policy once a compromised site is sanitized and no longer part of the botnet. Recall that each compromised site is cross linked to other compromised sites to increase search result ranking (Section 5.2). Therefore, when a site is no longer compromised, there is no value for the site to receive backlinks. Assuming the botmaster is actively monitoring the sites in the botnet, he should be able to adjust the cross linking policy to only link to sites that are still part of the botnet.

Using the set of sanitized sites described in Section 5.1.2, we track the number of backlinks received by each site over time from other compromised sites, noting whether a sanitized site still receives backlinks and for how long. In addition, we measure the average number of backlinks received before a site is sanitized, and after, to see whether the botmaster updates the cross linking policy to decrease the number of backlinks given to sanitized sites. Surprisingly, sanitized sites still overwhelmingly receive backlinks, and do so for long periods of time. Out of 508 sanitized sites, nearly all sites still receive backlinks even after being sanitized: all but two sanitized sites receive backlinks through February 26, and 488 (96%) through March 2.

In summary, it appears that the botmaster exerts only limited control over many compromised sites, letting many degrade over time. Further, this is but one of the inefficiencies in how the botnet is operated. While we do not have insight into the reasons for these lapses—whether negligence, lack of insight, or lack of need—the large numbers of stragglers and useless cross linking to sanitized sites makes it clear that in its existing regime the botnet does not reach its full potential impact.

## 5.2 Cross Linking

Next we examine the characteristics of the cross linking approach used by the SEO campaign to poison search results. Link "juice" is the SEO vernacular [16] for the number of back links received from other unique Web sites, a well-known feature used by search algorithms when ranking Web pages [18]. Consequentially, one of the primary requirements for the SEO campaign to effectively poison search results is to artificially accumulate juice. Thus, by understanding the campaign's cross linking strategy, we are in a better position to counter search result poisoning.

The SEO botnet performs link farming where, using the terminology of [7], a small subset of compromised Web sites emulate *authorities* and receives substantially more juice than the other sites emulating *hubs*. This relationship lasts for an extended time period and ends when the botmaster rotates authority sites, with a different subset of compromised sites becoming authorities and receiving the dominant fraction of juice, and previous authorities becoming hubs. Link farming benefits the botmaster in a couple of ways. First, because there is a non-linear relationship between search result position and the amount of traffic clicking through the search result, the botmaster can attract more traffic by focusing on having the authority sites occupy a handful of top search result positions, rather than many low search positions using all compromised sites. Second, by selectively "juicing" a relatively small subset of authorities, the botmaster can limit the number of sites lost due to interventions by the site owner or defense mechanisms like Google Safe Browsing.

In our study, we identified two major authority rotations by monitoring when the amount of juice received by sites from the botnet changes substantially. Both occur in conjunction with major changes in the botnet. The first rotation occurs on November 1, 2011, when the SEO kit was updated from OEM to MAC. The second rotation occurs on January 28, 2012, when the botnet directory server experienced an outage (Section 5.1.1). In both cases, it seems the botmaster initiated the rotations because they appear related to version changes on the control server.

Table 3 summarizes the distribution of "juice" among compromised nodes in the botnet. It shows the number of compromised sites, grouped by order of magnitude of the average daily back links received by each site, for each time period. For example, in the first period from the beginning of the study to November 1st, 2011, there are 532 nodes that receive less than ten back links, 71 nodes that receive 10–100 back links, and 12 nodes that receive 100–1000 back links. Each time range has a consistent pattern: a small subset of sites (authorities) receive hundreds of back links, tens of sites receive tens of back links, and the remaining hubs receive less than 10 back links. We confirmed that the actual roles of compromised sites indeed changed from one period to another. For example, in the Nov 2011 rotation over 80 nodes had their juice reduced by at least an order of magnitude, while 20 nodes had their juice increased by at

| Group | >03/06 |
|---|---|
| <10 | 665 |
| 10 – 100 | 250 |
| 100 – 1000 | 1 |
| >1000 | 63 |

**Table 4: The number of compromised sites grouped by the total amount of juice received from blog posts, after the release of v8.**

least an order of magnitude.

The top sites receiving the most juice are the most valuable to the botmaster since they have the most exposure in search results and attract the most traffic. At the same time, because they have the most exposure they are also the most vulnerable. To undermine SEO botnets, targeting these sites first will be most effective, e.g., via blacklists like Google Safe Browsing.

As noted in Section 3, the release of v8 introduces a new cross linking mechanism that uses blogspot blogs as an extra layer of redirection. We do not directly compare the amount of juice observed using this new mechanism with the botmaster's previous approach because there are only two posts per blog, the last of which occurred in early April. Since this strategy rotates juice in sporadic large batches, rather than periodic increments, we focus the v8 cross linking analysis to data after March 6, 2012. As with the previous link farming strategy, though, we see a similar distribution of sites that emulate authorities (63) and hubs (665), albeit with a larger number of middle-sized hubs (251) as shown in Table 4.

## 5.3 SEO Effectiveness

Using the earlier data of the Dagger cloaking crawler, we next examine the ability of the SEO botnet to poison search results in Google. These poisoned search results represent doorways, which redirect users who click through the search result, leading to a destination of the botmaster's choosing. Thus, the doorways accumulate traffic for the botmaster to monetize. Therefore, we assess the potential threat posed by the SEO botnet by measuring the poisoned search results as seen from the user's perspective.

We find that the botnet can be quite effective in bursts, with thousands of poisoned search results targeting popular, trending search terms at any given time during the burst. At its peak, it is the dominant contributor to poisoned search results in Google from April through June 2011. The botnet is able to juice specific search terms with poisoned search results in Google within 24 hours, whereas it takes Google over 48 hours to start to counteract the poisoning.

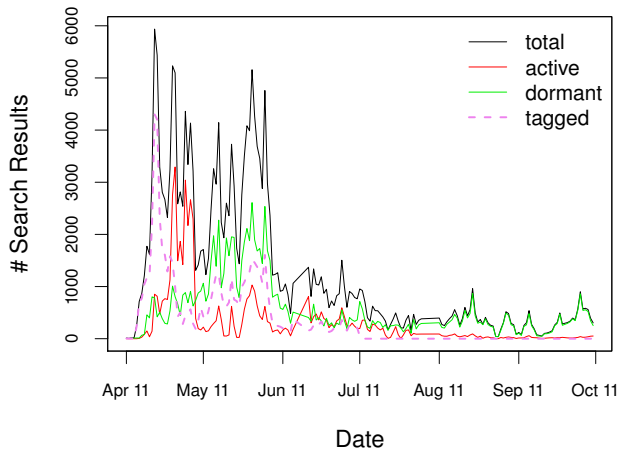


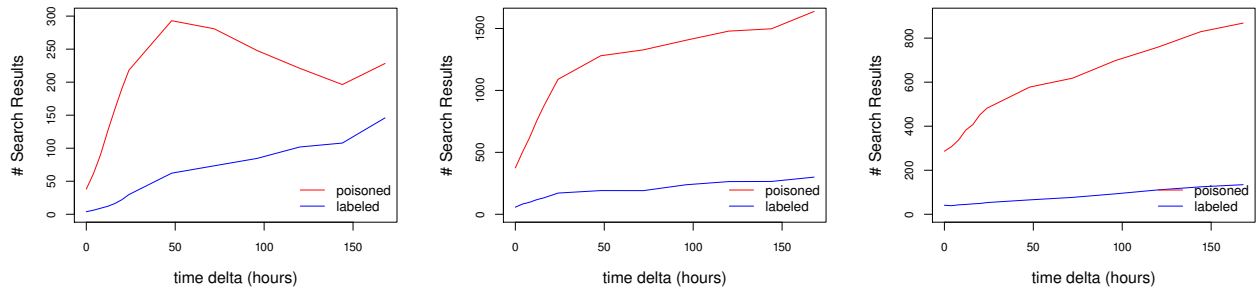

**Figure 5: Quantity of poisoned search results attributable to the SEO campaign. Each line shows which poisoned results are redirecting users, dormant, or tagged by Google Safe Browsing.**

### 5.3.1 Quantity

Ultimately, the goal of the SEO botnet is to attract user traffic to its compromised sites by manipulating search results. We first evaluate the effectiveness of the SEO botnet in achieving this goal by analyzing the placement of its sites in search results.

Using the set of compromised sites enumerated by the host crawler, we identify the botnet's poisoned search results using the URLs that link to a compromised site. We then characterize each poisoned search result into one of three states over time: active, tagged, or dormant. *Active* poisoned search results are cloaking and actively redirecting users. Users who click on these search results will be taken to an unexpected site, such as fake AV, to monetize their clicks. *Tagged* results have been labeled as malicious by Google Safe Browsing (GSB) [3], presumably discouraging users from visiting and preventing the botmaster from significantly monetizing traffic to these URLs. GSB blacklists URLs that lead to phishing and malware pages. Although not all pages the botnet uses to monetize traffic may fall under GSB's purview, when GSB labels those pages that do it is a useful indicator of defenses undermining the botmaster's SEO campaign. *Dormant* poisoned search results are cloaking but not redirecting. These search results lead to sites that apparently have redirection disabled, and the botmaster no longer derives value from them.

Figure 5 shows the evolution of an SEO campaign over time as viewed by the prevalence of the various kinds of poisoned search results. Over six months, we saw four main periods of activity. In a starting period, from April 1st to April 18th, most poisoned search results were tagged yet the volume of active remained high. On April 15th, for example, we observed 2,807 poisoned search results, of which 1,702 search results were tagged, 721 were active, and 384




(a) Start (b) Surge (c) Steady

**Figure 6: The number of poisoned search results attributable to the SEO campaign, when the same query is retried after a time delta. The POISONED line represents poisoned search results that have not been labeled by GSB, whereas the LABELED line represents poisoned search results that have been labeled by GSB.**

were dormant. The tagged and dormant search results are the remnants of the previous SEO campaign by this botnet, while the growing number of active results reflects increasing momentum of a new campaign.

The start period transitioned into a *surge* period from April 18th to April 28th, where a surge in active poisoned search results corresponds with a decline in tagged. This surge reflects the campaign promoting a new set of heavily "juiced" hub sites (Section 5.2). This 10-day window shows the botnet SEO campaign at its peak, with most poisoned search results actively monetizing traffic before Google Safe Browsing and site owners can react.

A third *steady* period, from April 28th to June 30th, exhibits a substantial decrease in active poisoned results and a corresponding increase in tagged and dormant results. These results are no longer effective for the botmaster since they have either been flagged by GSB to warn away uses or the sites have been sanitized by their owners.

After June is an *idle* period where the total volume of poisoned search results declines substantially, coinciding with the timeframe of an organized intervention into the fake AV industry by the FBI [8]. From July through October 2011 the SEO campaign had results linked to compromised sites that remain tagged or dormant, but only a negligible number of sites were actively redirecting. It highlights the successful impact of interventions that undermine the vector by which the botmaster could monetize traffic, like the fake AV takedown. Undermining monetization removes the key incentive for the botmaster to keep SEO campaigns active.

### 5.3.2 Temporal

Section 5.3.1 assesses the SEO campaign's activity level over time. However, quantity alone does not give a complete portrayal of the volume of poisoned search results and their coverage in spamming specific search terms. For example, on April 23, 2011, Dagger queried "the ten commandments list" and found one poisoned search result from this SEO campaign. Then, 16 hours after the initial query, Dagger re-queried "the ten commandments list" and found 20 poisoned search results. We suspect the increase in poisoned search results is due to the increased time available for the campaign to SEO their sites. Regardless, these dynamics demonstrate the importance of the time component in conveying impact. Thus, to assess the botnet's potential threat through volume and coverage of specific search terms, we also measure the quantity of poisoned search results for the same query at subsequent points in time.

Recall that Dagger repeatedly queries for the same search terms over time to enable precisely these kinds of temporal analyses (Section 4.2). Figure 6 presents the number of poisoned search results attributable to the SEO campaign when the same query is repeated for varying time deltas. For each search term, the zero time delta is when Dagger first sees poisoned search results for that search term. We show results separately for the start, surge, and steady periods to highlight differences among them.

Each graph contains two lines. The POISONED line represents the number of poisoned search results that have not been labeled by Google Safe Browsing, averaged across the entire data set. We use the same methodology as above, except here we apply it to the additional temporal data and we do not distinguish between active and dormant poisoned search results. Conversely, the LABELED line represents the average number of poisoned search results that have been labeled by GSB. Not surprisingly, we see the same results as before. The start period has a mixture of poisoned search results and labeled search results from a previous SEO campaign. Then there is a burst of poisoned search results during the surge period, and a steady stream of poisoned search results in the steady period.

The number of poisoned search results from their first appearance in a query for a search term is just the tip of the iceberg. In other words, within hours of the initial appearance, users are likely to encounter a flood of poisoned search results. Although applicable for all time periods, it is most prominent during the surge period as the number of poisoned search results seen increases nearly 3× from 374

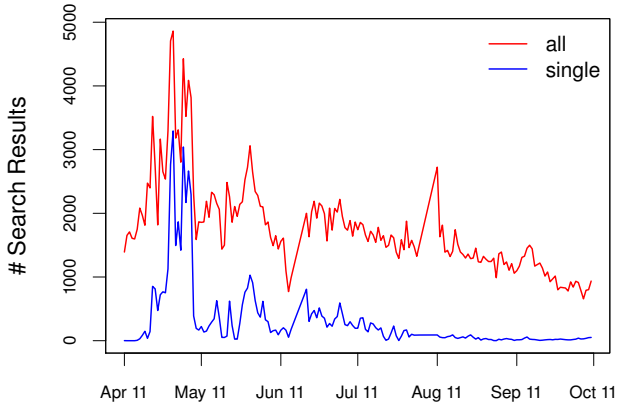



**Figure 7: Comparison between this SEO campaign against all actively redirecting poisoned search results.**

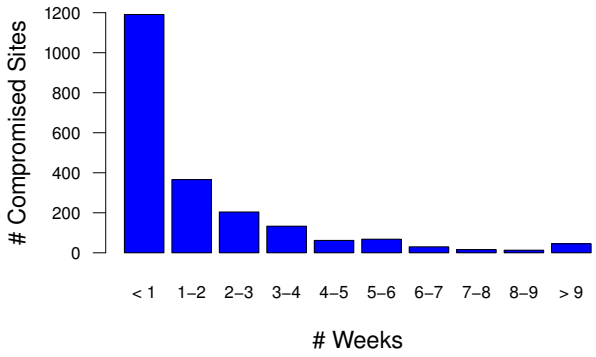



**Figure 8: Duration of compromised sites in poisoned search results that actively redirect users to scams.**

to 1,089 within 24 hours. In the start period, we see an increase from 38 to 218 within 24 hours, and in the steady period we see an increase from 286 to 482 within 24 hours.

Further, at the start of a new campaign GSB lags the increase in poisoned search results that arrive shortly after the initial appearance: the slope of the POISONED lines is higher than the LABELED lines during the surge and steady period. Only when the campaign enters the start period does GSB begin to react swiftly to the later arriving poisoned search results (the dip after 48 hours in the POISONED line).

#### 5.3.3 Market Share

Intersecting the data from the botnet crawler and the search crawler also allows us to compare the relative size of this SEO campaign against all other similar SEO campaigns in terms of the volume of poisoned search results. Figure 7 shows the number of cloaked search results over time found by Dagger, and then the subset attributed to just this SEO campaign. During the surge period, this campaign accounted for the majority of cloaked search results at 58%. This campaign was most prominent on April 24th, when 3,041 out of 4,426 (69%) poisoned search results came from this single campaign. Even as the surge decreased in the steady period, the SEO campaign still accounted for 17% of all active poisoned search results observed. As a result, not only is this SEO campaign one of the main contributors of poisoned search results, it has demonstrated the potential to poison more search results than all other competing SEO campaigns combined for an extended duration of 10 days.

#### 5.3.4 Active SEO Duration

Similar to Section 5.1.2, we next quantify how long the botmaster is able to effectively utilize compromised sites as doorways for funneling users to scams. Compared to lifetime, which measures how long a site remains part of the botnet, here we focus on active duration, the total amount of time that the site is both exposed to users through poisoned search results and actively redirects them to scams.

For each compromised site we collect all occurrences of poisoned search results to the site observed from April 1 to September 30, 2011. In addition, we track whether the poisoned search results were redirecting users and whether they were labeled by GSB. We use the first occurrence of a poisoned search result to a site as the start of the site's active duration. We end the active duration when we do not see another occurrence of a poisoned search result, attributable to this site, within the following three weeks of the last result.

In this six-month period, 3,822 compromised sites from this campaign were involved in poisoning search results. Of these, 2,128 sites (56%) actively redirected users. Figure 8 shows a binned histogram of the number of sites that poison search results and actively redirect users at a week granularity. A majority of actively redirecting sites (56%) have durations of less than a week. Of the remaining 937 sites that effectively poison search results and redirect users for longer than a week, most (89%) survive 1–6 weeks.

### 5.4 Monetization

Ultimately, the botmaster operates and maintains the GR botnet to make money. Compromising sites, cloaking and traffic segmentation, cross linking and poisoning search engine results, etc., are all component parts of a black hat SEO machine engineered to make a profit.

#### 5.4.1 Scams Targeted

Using the data from the Trajectory redirection crawler, we categorize the redirection chains that lead users from poisoned search results to the different scams used by the GR botnet to monetize traffic. Specifically, we selected redirection chains that: (1) originated from one of the doorway pages, (2) contained more than one cross

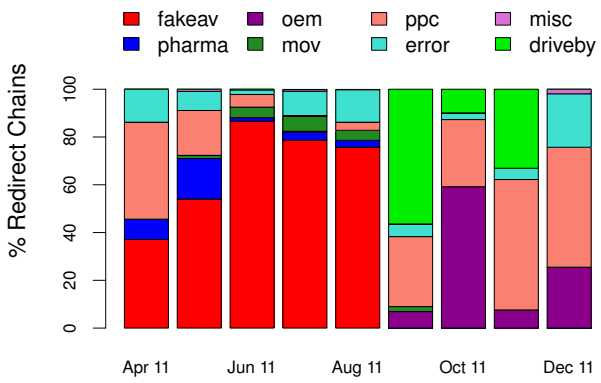



**Figure 9: Relative breakdown of the categories of scams that poisoned search results ultimately take users.**

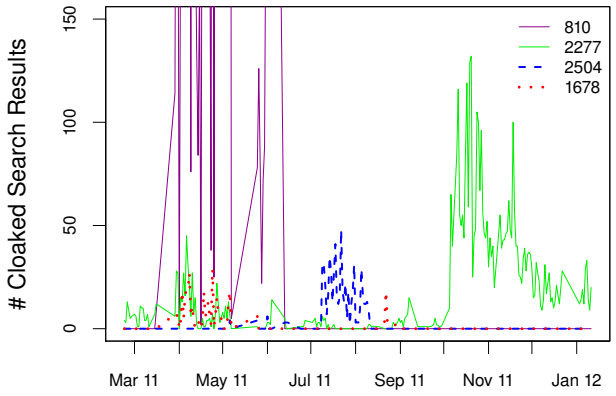



**Figure 10: Number of poisoned search results that lead to RivaClick over time. Each line represents a unique affiliate ID. The y-axis is truncated at 150 to show details (the max y-value is 1,231).**

site redirection, and (3) occurred while mimicking a Microsoft Windows user running Internet Explorer; as discussed below in Section 5.4.2, the majority of redirect chains observed while mimicking a non-Windows user generally led to the RivaClick pay-per-click affiliate program. We manually clustered the redirection URLs based on similar URL characteristics, such as the same PHP file with the same HTTP GET parameters and arguments. For example, although `http://model-seil.ru/afro/index.php` and `http://softwarename.ru/protect/index.php` appear to represent two separate hosts, they in fact resolve to the same IP address. After clustering, we constructed a network graph starting from doorways and ending at different kinds of scams. This graph allows us to trace the scams where the botmaster was an affiliate.

Previous work noted that SEO campaigns in general shift between different affiliate programs over time [19]. Therefore, for this analysis we arbitrarily divided the redirect chains by the month when the chain was observed. Figure 9 shows the relative breakdown of the kinds of scams that the redirection chains take users. The "misc" category refers to crawls that we could not classify, such as redirections that ultimately led to the Google search page, and "error" are crawls that returned an HTTP error code or screenshot.

We see two distinct periods of scam targeting, with the transition between the two coinciding with the 2011 fake AV takedown [8]. Early on, from April through August, the botnet redirects the majority of poisoned search results to fake AV programs, presumably because of their profitability [17]. We also see a varying amount of redirection chains leading to counterfeit pharmaceutical programs, including the GlavMed, Mailien, and RX-Partners programs, although not nearly as prevalent as fake AV. From June through August, we also see an increase in the proportion of search results directed to `movdl.com`, a pirated media affiliate program. Redirection chains to `movdl.com` stop in September, though.

After the fake AV takedown, the botmaster markedly changes the scams targeted. In September, we see the intermediary node that sent traffic to the one remaining fake AV program now sending traffic to a drive-by download affiliate program. This target is also temporary, as by October the botnet updates the SEO kit to version OEM and redirects the majority of the traffic to OEM affiliate programs (TheSoftWareSellers and OEMPays), which continues until December when we found that the GR botnet stops redirecting. Finally, pay-per-click is notably a steady safety net throughout, and we explore it in more detail next.

### 5.4.2 RivaClick Traffic Affiliate Program

Recall from Section 3.3 that we downloaded past versions of the SEO kit. While crawling these past versions, we found that the SEO campaign was actively redirecting users to the URL:

```
http://www.rivasearchpage.com/
?aid=2277&said=0&n=10&q=[query]
```

This URL leads to a feed of URLs for the RivaClick Traffic Affiliate Program. RivaClick operates similarly to other Internet advertising platforms. There are advertisers who want to buy traffic for a specific topic of interest, usually determined by the user's query string, and there are publishers who sell traffic. RivaClick groups the advertisers' links into a feed, which is provided to publishers who will receive commissions on click traffic to links from the feed. An important difference between RivaClick and other advertising platforms is that RivaClick provides little guarantees about

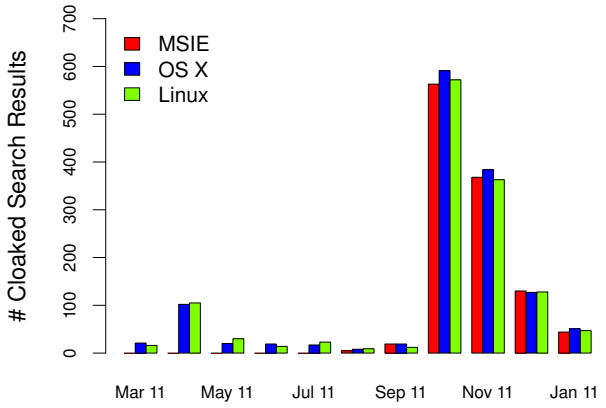



**Figure 11: Number of poisoned search results that lead to RivaClick depending on the OS/browser.**

the quality of the traffic being sold, which allows publishers to dump traffic obtained through search result poisoning. Based on the URL extracted from a previous SEO kit and the HTTP GET parameters from the URL, it appears that the botmaster is an affiliate of RivaClick with ID 2277.

With this affiliate identifier in hand, we retroactively examined poisoned search results from the Trajectory crawler starting in March 2011. One pass of the search crawler captures the entire redirect chain for poisoned search results, from the doorway, returned when the user first clicks the search result, through all the intermediary hops, and finally the landing page. In this section, we focus on redirect chains that landed on RivaClick.

Figure 10 shows the quantity of poisoned search results funneled into RivaClick per day for the four most-frequently seen affiliates: 810, 1678, 2277, and 2504. Because we found these affiliates performing search result poisoning, we assume they are running similar black hat SEO campaigns. Therefore, we compare the four affiliates to provide a sense for the relative size of the GR botnet and its peers and competitors in terms of the number of search results leading to RivaClick (a more focused comparison than Figure 7, which is in terms of all poisoned search results).

The GR botnet, as affiliate 2277, redirected a small but steady number of search results to RivaClick for much of 2011, but then significantly increases results to RivaClick starting in October 2011 after the fake AV takedown. Meanwhile, the other affiliates were burstier. 1678 directed a small burst from April–May, and 2504 directed a burst from July–August. Finally, 810 redirected bursts from March–June, but with far more intensity (max of 1231 on June 2). As a result, it appears that 2277 is a long lasting, relatively mid-size SEO affiliate of RivaClick.

Figure 11 focuses more closely on affiliate 2277. The Trajectory search crawler visited each poisoned search result while mimicking three different browsers: Microsoft Internet Explorer running on Windows, Mozilla Firefox running on Mac OS X, and Mozilla Firefox running on Linux. These three visits enable us to analyze the traffic segmentation policy employed by the botmaster based on browser and operating system. Indeed, it appears that such demultiplexing occurred from March through September. As seen in Figure 11, only Mac OS X and Linux traffic led to RivaClick. Starting in August, when the botmaster could no longer monetize Windows traffic through fake AV scams, traffic from all platforms were redirected.

## 6 Conclusion

Overall, we find that with modest resources the GR botnet can be very effective in poisoning search results, becoming for months at a time the dominant source of poisoned results. At the same time, we have seen two kinds of intervention against the SEO botnet. The first targets the botnet directly, its infrastructure (compromised sites) and vector (poisoned search results). Given that sites remain compromised for months, cleaning up sites has not been effective at undermining the botnet; indeed, even when we explicitly notified site owners about the malware, few reacted or responded. Google, however, is more responsive, tagging poisoned search results within a couple of days—but that window is still presumably effective for the botmaster given the intensity of the SEO activity. The second undermines monetization, and appears to be much more effective. With evidence of the importance of a "killer scam" in monetizing and driving innovation in SEO campaigns, we observe substantially more activity from the botnet when the fake anti-virus market is stable, whereas the botmaster appears to scramble to monetize traffic when the fake anti-virus market is in flux and the GR botnet becomes relatively idle. Undermining monetization appears to be a potent response to these types of attacks.

## Acknowledgments

We thank Damon McCoy for insightful comments and discussion of this work, Neha Chachra for the Trajectory crawler, Erin Kenneally for legal oversight, the anonymous reviewers for their valuable feedback. We also thank David Dagon, the Georgia Tech Information Security Center, and Damballa for domain sinkholes. This work was supported in part by National Science Foundation grants NSF-0433668 and NSF-1237264, by the Office of Naval Research MURI grant N000140911081, and by generous support from Google, Microsoft, Yahoo, and the UCSD Center for Networked Systems (CNS).

## References


[1] eMarketer. eMarkter Press Release: Google's Share of US Search Revenues is Still Growing. http://www.emarketer.com/PressRelease.aspx?R=1008258, Mar. 2011.

[2] D. Goodwin. Search Ad Spending Could Hit $19.51 Billion in 2012. http://searchenginewatch.com/article/2143093/Search-Ad-Spending-Could-Hit-19.51-Billion-in-2012-Report, Feb. 2012.

[3] Google. Google Safe Browsing API. http://code.google.com/apis/safebrowsing/.

[4] C. Grier, L. Ballard, J. Caballero, N. Chachra, C. J. Dietrich, K. Levchenko, P. Mavrommatis, D. McCoy, A. Nappa, A. Pitsillidis, N. Provos, Z. Rafique, M. A. Rajab, C. Rossow, K. Thomas, V. Paxson, S. Savage, and G. M. Voelker. Browser Exploits as a Service: The Monetization of Driveby Downloads. In *Proceedings of The 19th ACM Conference on Computer and Communications Security*, October 2012.

[5] J. P. John, F. Yu, Y. Xie, A. Krishnamurthy, and M. Abadi. deSEO: Combating Search-Result Poisoning. In *Proceedings of the 20th USENIX Security Symposium*, August 2011.

[6] C. Kanich, C. Kreibich, K. Levchenko, B. Enright, V. Paxson, G. M. Voelker, and S. Savage. Spamalytics: an Empirical Analysis of Spam Marketing Conversion. In *Proceedings of the ACM Conference on Computer and Communications Security*, Oct. 2008.

[7] J. M. Kleinberg. Authoritative Sources in a Hyperlinked Environment. In *Proceedings of the 9th ACM-SIAM Symposium on Discrete Algorithms*, 1998.

[8] B. Krebs. Huge Decline in Fake AV Following Credit Card Processing Shakeup. http://krebsonsecurity.com/2011/08/huge-decline-in-fake-av-following-credit-card-processing-shakeup/, August 2011.

[9] C. Larsen. Latest SEP (Search Engine Poisoning) Research, Part 1-7. http://www.bluecoat.com/security/security-archive/2012-02-15/latest-sep-search-engine-poisoning-research-part-1, Feb. 2012.

[10] N. Leontiadis, T. Moore, and N. Christin. Measuring and Analyzing Search-Redirection Attacks in the Illicit Online Prescription Drug Trade. In *Proceedings of the 20th USENIX Security Symposium*, August 2011.

[11] K. Levchenko, N. Chachra, B. Enright, M. Félegyházi, C. Grier, T. Halvorson, C. Kanich, C. Kreibich, H. Liu, D. McCoy, A. Pitsillidis, N. Weaver, V. Paxson, G. M. Voelker, and S. Savage. Click Trajectories: End-to-End Analysis of the Spam Value Chain. In *Proceedings of the IEEE Symposium and Security and Privacy*, Oakland, CA, May 2011.

[12] L. Lu, R. Perdisci, and W. Lee. SURF: Detecting and Measuring Search Poisoning. In *Proc. of The 18th ACM Conference on Computer and Communications Security*, October 2011.

[13] M. Maunder. Zero Day Vulnerability in many WordPress Themes. http://markmaunder.com/2011/08/01/zero-day-vulnerability-in-many-wordpress-themes/.

[14] D. McCoy, A. Pitsillidis, G. Jordan, N. Weaver, C. Kreibich, B. Krebs, G. M. Voelker, S. Savage, and K. Levchenko. PharmaLeaks: Understanding the Business of Online Pharmaceutical Affiliate Programs. In *Proceedings of the 21th USENIX Security Symposium*, 2012.

[15] T. Moore, N. Leontiadis, and N. Christin. Fashion Crimes: Trending-Term Exploitation on the Web. In *Proceedings of The 18th ACM Conference on Computer and Communications Security*, October 2011.

[16] SEOmoz. PageRank, Link Patterns & the New Flow of Link Juice. http://www.seomoz.org/blog/pagerank-link-patterns-the-new-flow-of-link-juice, May 2007.

[17] B. Stone-Gross, R. Abman, R. Kemmerer, C. Kruegel, D. Steigerwald, and G. Vigna. The Underground Economy of Fake Antivirus Software. In *Proc. of the 10th Workshop on the Economics of Information Security (WEIS)*, 2011.

[18] A.-J. Su, Y. C. Hu, A. Kuzmanovic, and C.-K. Koh. How to Improve Your Google Ranking: Myths and Reality. In *Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence*, August 2010.

[19] D. Y. Wang, S. Savage, and G. M. Voelker. Cloak and Dagger: Dynamics of Web Search Cloaking. In *Proceedings of The 18th ACM Conference on Computer and Communications Security*, October 2011.

[20] Y.-M. Wang, M. Ma, Y. Niu, and H. Chen. Spam Double-Funnel: Connecting Web Spammers with Advertisers. In *Proceedings of the 16th International World Wide Web Conference (WWW'07)*, pages 291–300, May 2007.

[21] B. Wu and B. D. Davison. Cloaking and Redirection: A Preliminary Study. In *Proc. of the SIGIR Workshop on Adversarial Information Retrieval on the Web*, May 2005.