Poisoned search results: How hackers have automated search engine poisoning attacks to distribute malware.

Fraser Howard & Onur Komili

SophosLabs

fraser.howard@sophos.com, onur.komili@sophos.com

Executive Summary
This paper describes recent research by SophosLabs into how attackers are using blackhat Search Engine Optimisation (SEO) techniques to stuff legitimate websites with content designed to rank highly in search engine results, yet redirect users to malicious sites. These websites are being used to distribute rogue security products (also known as "scareware" or "fake anti-virus") onto users' computers.

Sophos researchers have analysed the malicious SEO kits used by hackers to create networks of thousands of cross-linked pages containing search-friendly content on hot-trending topics, hosted on compromised, legitimate websites.

1 Introduction

Search engine optimisation (SEO) is a generic term given to a range of tricks and techniques that are used in order to elevate the ranking of a particular URL in the results listings of search engines. Successfully done, SEO can have a significant effect upon the volume of traffic hitting a site. Keen to boost their search engine ranking, many organisations will recruit marketing consultants to to optimise their site content for search engine indexing. The major search engines publish guidelines on how to improve results rankings [1,2,3]. At the opposite end of the spectrum a range of techniques may be used to achieve the same boost, but in an unscrupulous way. The term *black hat SEO* is often applied to this latter case [4,5].

Historically, black hat SEO is something we might have associated with spammers, scammers and disreputable on-line merchants. More recently however, it is being used to drive user traffic to malicious sites for the purposes of distributing malware, particularly fake anti-virus Trojans (often termed *scareware* [6]).

At the time of writing we see numerous examples of these attacks hitting the press each day, with security firms and journalists reporting how searching the Internet for the latest topical news item can lead to infection [7,8,9]. In this technical paper, we provide details from the analysis of some of the kits that are being actively used in SEO attacks. Through this we hope to provide insight into how these attacks are being managed, and how to best defend against them.

A brief definition of some of the nomenclature that is repeatedly used within this paper in discussing SEO attacks is provided below:

- *Fake anti-virus* – class of malware that inundates the user with fake security alerts in order to trick them into paying to register the rogue security product.
- *SEO page* – the keyword-stuffed pages designed to rank highly in search engine results, yet redirect users to rogue sites. Sometimes called *SEO poisoned pages*.
- *SEO kit* – the application used to create and manage an SEO attack site. Responsible for generating SEO pages for search engine crawlers which poison search results in order to redirect users to rogue sites. Sometimes called *blackhat SEO kits* (but within the context of this paper we are referring specifically to kits focussed on malware distribution).
- *SEO poisoning* – a term used to describe the process of tricking the search engines into ranking an SEO page high up in the search results. Those results can be regarded as "poisoned".
- *Search engine crawler* – another term for a web *bot* or *spider*, which refers to a computer program that browses the web in a structured fashion in order to index pages and collect data that can be readily searched.

2 Overview of SEO attacks

In concept, SEO driven attacks are pretty simple. The attackers use SEO kits (PHP scripts typically) to create web pages stuffed with topical keywords and phrases that will be consumed by search

engine crawlers. Then, when a user searches for such keywords they are presented with a link to the SEO page high up in the search engine results. Clicking on the link is all it takes for the user to be exposed to malware. The SEO kit recognises they have arrived via a search engine and redirects them to some malicious site. This is illustrated in Figure 1.

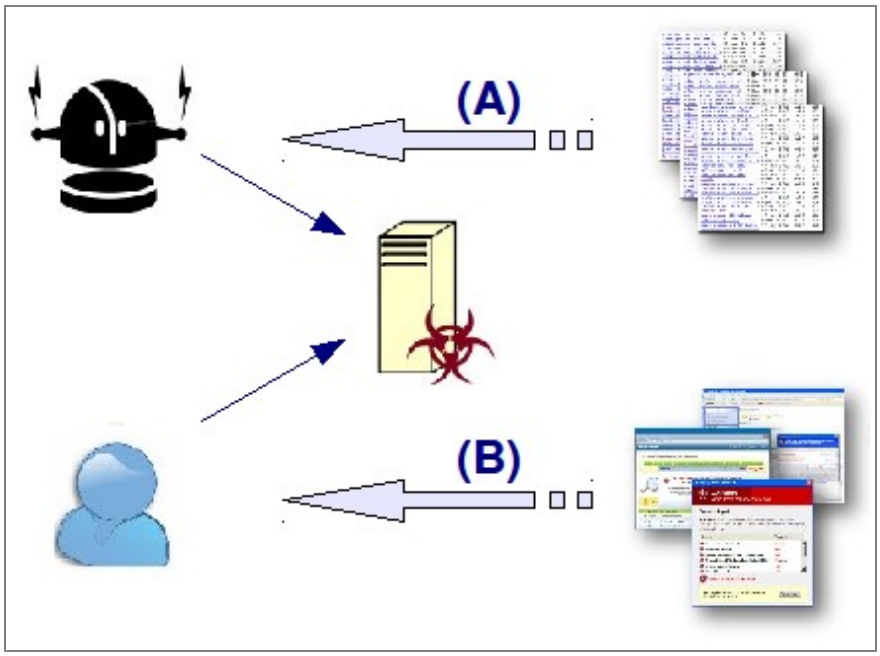

Figure 1: Overview of how SEO driven attacks work. The core SEO kit (hosted on a compromised legitimate site) uses scripting to feed search engine crawlers keyword-stuffed pages (A), but redirect users that have arrived via search engines to malicious sites (B).

Once redirected from the SEO page, there may be multiple additional levels of redirection before the final payload is actually delivered. For example, in the current SEO attacks being used to distribute fake anti-virus malware, the victim is typically redirected at least twice before being presented with the fake anti-virus web page (which tricks them into believing their system is infected, and installing the malware that masquerades as a security product).

3 Hosting the SEO kits

Numerous *SophosLabs* blog posts have described how it has become almost routine for attackers to compromise legitimate web content in order to distribute malware [10,11,12]. In fact, once a site is compromised it can be abused in a whole variety of ways – from hosting phishing sites [13] to providing a platform from which other attacks can be performed. It is no surprise that compromised sites are also being used to host current SEO attacks.

The reason for this is not just about traceability or abusing someone else's resources (hosting, bandwidth etc.). By hosting the SEO attack within a legitimate site, the attackers are able to piggyback on the reputation of that site, making it harder for the search engines to identify and remove the rogue links. Additionally, distributing attacks across multiple compromised host sites provides increased resilience against URL filtering and other defensive mechanisms.

All except one of the SEO attacks investigated within this research were hosted within legitimate sites. It is hard to establish exactly how these sites had been compromised without access to

additional log data. However, there was often a common link found between the compromised sites. For example, the use of the same Content Management System (CMS), including (but not limited to) *Joomla!*, *Wordpress*, *phpBB, MediaWiki, osCommerce, CMS Made Simple* and *Zen cart*. This suggests that it is vulnerabilities in the CMS (or CMS plug-ins and extensions) that are being exploited by attackers in order to compromise the sites. For one attack, we discovered that all of the compromised sites involved were hosted by the same provider, suggesting that they had been compromised via server vulnerabilities.

The one attack where the SEO pages were not hosted within the legitimate site proved to be an interesting case. We had identified SEO pages being hosted on sub-domains, where the sub-domain consisted of a string based on the keywords (for example *ford-police-interceptor.<domain>.com* or *sandra-bullock-divorce.<domain>.com*). Further investigation suggested it was actually the hosting provider who had been compromised, such that sub-domains for all the domains they hosted resolved to a malicious IP, on which the SEO kit was hosted.

URL	IP	Comments
www.h********a.com	216.***.***.40	Legitimate hosting provider, located in North America.
seo-keywords.h********a.com	212.***.***.139	Suspicious IP, located in Luxembourg

Table 1: IP addresses for a legitimate site and its SEO-related sub-domain illustrating how the sub-domain resolves to a rogue IP. (Domain names and IP addresses partially obscured.)

The kit would generate an SEO page based on the string provided as a sub-domain that was hosted on any domain on this provider. If a *major* hosting provider was ever compromised in a similar fashion, there could be thousands of domains spewing SEO poisoned content in that single attack. At the same time since it is a single point of entry, the attack can be quickly stopped once identified.

In one of the attacks investigated, the SEO kit was always accompanied by a PHP backdoor component as well, providing the attacker with the ability to issue system commands, upload and download files and launch remote attacks.

4 Managing the SEO attacks

Once the host site has been identified and successfully compromised, the SEO kit will be uploaded and installed. In this section of the paper we will present an overview of the functionality in the SEO kits that have been seen thus far.

4.1 Branching logic: user or search engine?

At the heart of the SEO attack is the ability to feed search engine crawlers content to index and redirect users to malicious sites. To achieve this it is necessary to distinguish between the various origins for the page request:

- search engine crawlers
- users arriving via search engines

- users just happening upon the page

The kits analysed typically used a central PHP script to handle all page requests. This makes it easy to make the above distinctions for incoming page requests. To identify crawlers, the IP of the originating page request can easily be queried (using `$_SERVER['REMOTE_ADDR']`). This can then be compared against the IP ranges typically used by the search engines (a check done by one of the kits we have analysed). An additional or alternative check may also be made on the user-agent string (using `$_SERVER['HTTP_USER_AGENT]`), which can be useful to flag requests from the crawlers that use distinguishing strings [14].

When the PHP script determines the page request is from a crawler, it will return the appropriate keyword-stuffed content. This may be generated dynamically (see section 4.2), or loaded from a file on disk (if a page relevant to the keywords being queried already exists).

Page requests from users arriving via search engines are identified by checking the referrer (using `$_SERVER['HTTP_REFERER']`) against a list of strings associated with the major search engines:

- aol
- bing
- google
- msn
- search
- slurp
- yahoo
- yandex

In one of the kits analysed the referrer check was even more rudimentary – simply inspecting the referrer URL for multiple occurrences of '&', which is the character normally used in the query string to delimit field-value pairs.

So it is trivial for the SEO kit to make the aforementioned distinctions for the origins of incoming SEO page requests. Of course, it is possible to deliberately remove or fake the HTTP headers associated with a request, or in fact, use a browser plug-in to do so [15,16]. But this is irrelevant, and not applicable to the vast majority of users.

It should also be mentioned that in at least one of the kits investigated, the IP for the originating page request was logged. The malicious redirect would only be delivered once to any single IP.

4.2 SEO page generation

For SEO driven attacks to succeed, the generated SEO pages have to satisfy search engine crawlers and get themselves ranked highly in the search results. Depending upon how the SEO kit works, the

links that the search engine index may be to either static HTML pages, the central PHP script or to SEO related sub-domains (see section 3). Examples of the first two are shown in Figure 2.

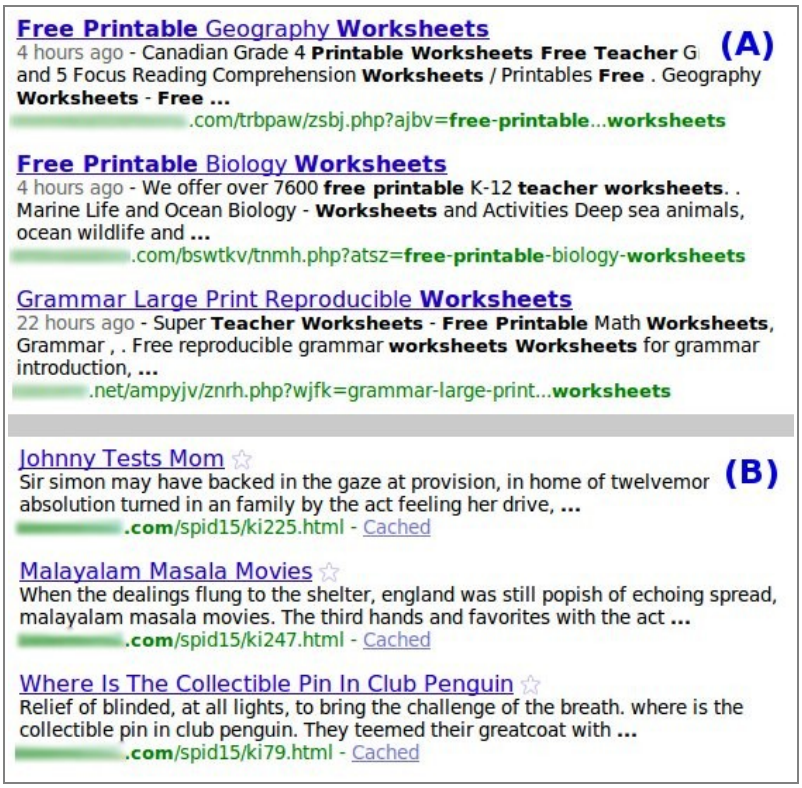

Figure 2: Example search engine listings showing links to SEO pages for kits using a PHP script to handle all requests (A) or static HTML pages (B). The relevant keywords are evident in the query string for (A).

Inspection of the source of the generated SEO pages reveals the expected mix of topical keywords and phrases together with links to other SEO pages on that site. The pages may appear ugly to the human eye, but this is immaterial – they only have to be "attractive" to search engine bots. For some of the kits analysed, the pages also contain links to SEO pages on other sites compromised with the same SEO kit. This practice is used frequently in SEO, and is known as *link exchange* [17].

Typically, the SEO pages are generated dynamically by the SEO kit, using search engines to source the relevant text for the page content. The kits use either the PHP client URL library, *cURL* [18], or `fsockopen()` [19] to retrieve the search results, from which the text can be extracted. We have seen both *Google* and *Bing* being used by the SEO kits analysed, but any of the major search engines could just as easily be used. Figure 3 illustrates the text that would be extracted from some search results by one of the kits analysed.

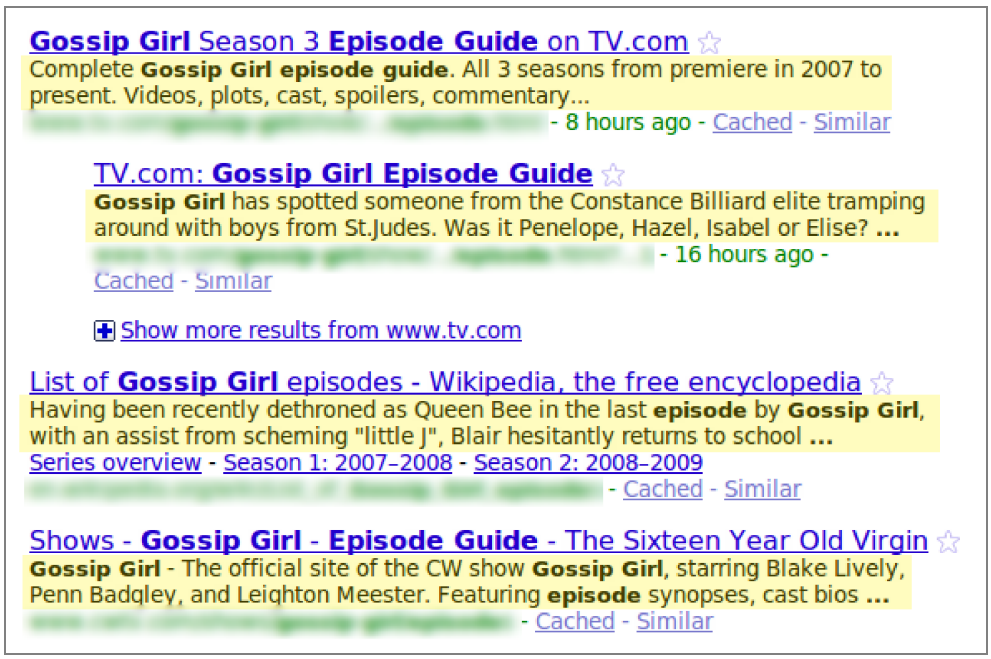

Figure 3: Illustration of the text extracted from search engine results by one of the SEO kits in generating the SEO page for related keywords.

The generated SEO pages contain links to other similarly generated pages, which requires the kit to maintain a list of URLs to other SEO pages (that are on the same site or on other sites compromised with the same SEO kit). For some kits this list is centralised and retrieved from a C&C server, in others it is maintained locally.

A summary of the steps typically involved in generating the SEO pages is listed below.

- Fetch latest keywords to poison from the C&C server or *Google Trends (*or similar). See Section 4.3.
- Download search engine results (using `fsockopen()`, *cURL*) for the relevant keywords.
- Use regular expressions to extract meta content from these results pages.
- Obtain list of links to other SEO pages on the same server, and on other servers compromised with the same SEO kit – see Section 4.5. (*Optional, not all kits include these links.)*
- Intersperse extracted meta content with the links to other SEO pages (to boost page rankings). Depending on the kit, other content may also be added as well, including:
  - random time/dates
  - images and video content

Once complete, the generated page is fed back in response to the HTTP request.

Some of the more sophisticated kits cache the generated content, storing a copy of each page generated for a particular set of keywords, as shown in see Figure 4. Aside from being more efficient, this also would reduce the possibility for the search engine providers to flag suspicious incoming queries (from a relatively small volume of IPs).

Figure 4: Screenshots of directory listings showing the cache used by two of the SEO kits analysed. On the left, content is cached using the keyword string as the filename. On the right, the filename is based on an MD5 of the keywords (lower-cased).

Inspection of such a cache reveals the search terms that users are querying before clicking through to the SEO site. Clusters of cached files appear for the popular search terms reflecting the slightly different keywords used in the search queries.

4.3 Selection of SEO terms to poison

For the SEO attack to succeed the keywords chosen are critical. The attacks rely on using content that users are actively seeking. The murky history of SEO contains abundant references to something known as *scraping* or *splogging*. This involves the copying of page content for the purpose of either driving traffic to a rogue site to profit through ads, or to promote linked affiliate sites.

As described above, the malware-oriented SEO attacks that we see today are similar in concept – scraping content from search engine results in order to generate the SEO pages. Much of the heavy lifting is already done by the search engines (and potentially other popular web applications), making it trivial to track hot topics [20,21] and extract related phrases (see Figure 5).

Figure 5: Extracts from the search results for 'NCAA Basketball tournament 2010 scores' from Bing (top) and Google (bottom) highlighting the ease with which related search terms can be automatically extracted.

The terms to poison may also be managed centrally at the C&C server, although this makes the attack less resilient against takedown efforts. In either case, a copy of the terms will often be stored locally (see Figure 6).

```
sextingjoslynjames@1268997144
espn ncaa basketball scores@1268997840
villanova university@1268998084
ohio universitw==@1268998804
kansas basketball@1269000058
google kalender@1269001020
acc tournament 2010 results@1269001727
god of war wiki@1269001736
uefa champions league draw@1269002352
tennessee basketball@1269002400
headley@1269002768
matt cooke@1269002831
obama fox news@1269003011
ncaa wrestling championships 2010 brackets@1269003481
seth binzer@1269003694
rio tinto@1269004157
quartet@1269005653
mutton@1269005999
butler university@1269006625
john sheehan@1269006638
stefani germanotta@1269006814
uefa champions league draw live@1269006887
rebecca gayheart@1269007168
health care reform bill summary@1269007610
fulham@1269007723
```

Figure 6: Snapshot of the list of search terms stored locally by one of the SEO kits analysed. This includes the time of when each term was first added (from Fri Mar 19th 11:12:24 2010 to Fri Mar 19th 14:08:43 2010 in this snapshot).

With one of the kits analysed we found a number of log files containing information about the page requests received, including:

- keyword(s) the user arrived by
- timestamp
- user agent string

- referrer string
- IP from which page request is received

The kit enabled the attackers to retrieve the log file via a HTTP request (using a specific $_GET parameter). This suggests that the attackers monitor the requests in order to more efficiently lure users in (for example by optimising keywords or search engines targeted).

Analysis of this log data reveals some interesting facts, some of which are highlighted in Table 2. It should be noted that this data is extracted from just one of the attacks investigated in this research. Caution should be applied before trying to use it to draw conclusions about the global state of SEO. Nonetheless, it is an interesting insight into one of the attacks that is currently active.

Category	Description
Top keywords	Over 950 unique keyword terms, dominated by terms associated with high profile celebrities and news: 1. michelle mcgee beauty 2. charles manson escape 3. nujabes dead 4. josyln james text messages 5. march madness on demand 6. espn brackets 7. americas next top model cycle 14 episode 2 8. michelle mcgee 9. watch justified online 10. dennis kucinich wife
Top IPs (for requests the SEO kit classified as from search engine crawlers)	Over 1500 unique IPs. 1. Google 2. Yahoo 3. Yandex 4. ScoutJet (crawler for new *blekko* search engine [22]) 5. Gigablast 6. Sogou
Top IPs (for requests via search engine referrers)	Over 3100 unique IPs, globally distributed (see Figure 7).
Top referrers	For this particular kit 98% of the referrers were Google, 0.9% Comcast and 0.4% Yahoo. 1. google.com (87%) 2. google.ca (2.5%) 3. google.co.uk (1.4%) 4. search.comcast.net (0.9%) 5. google.de (0.8%) 6. google.pl (0.7%) 7. google.com.au (0.6%) 8. google.com.br (0.5%) 9. google.co.in (0.4%) 10. google.com.mx (0.4%)

Table 2: Some of the data extracted from the page request logs for one of the SEO kits analysed. Covers the period March 18th-23rd, 2010.

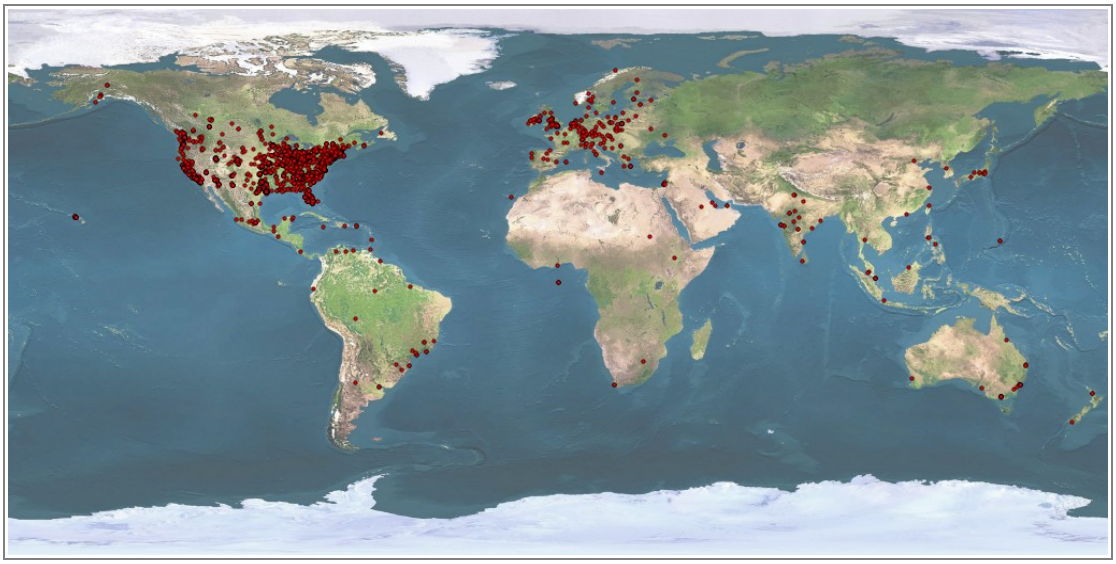

Figure 7: Global distribution of IPs logged as hitting one of the SEO attacks between March 18th and 23rd.

4.4 Redirection of user traffic

When incoming requests to the SEO pages are determined to be from users via a search engine, the requests are redirected to a malicious site. There are many ways of achieving this, the simplest of which is to use the PHP `header()` function to set the appropriate response-header field [23] and send a redirect (302) status code back to the user's browser.

Some of the kits produce the same page content for both search engine bots and users. In this scenario the redirection of users is achieved by active content embedded in the web page. The search engine bots will simply parse and index the raw page, most likely remaining oblivious to the redirect. Users on the other hand, will be redirected when the SEO page is loaded in their browser. For the more sophisticated of these kits, a check on the user-agent string is used to identify requests from search engine bots, such that the active content for redirection is only added for requests from users. This is presumably to help evade detection and blacklisting by the search engines.

There are many ways to redirect the user using active content within the web page, the most common being JavaScript or *ActionScript* in embedded Flash content. A couple of examples used in recent SEO attacks are shown in Figure 8.

Figure 8: Examples of redirection from within the SEO page. (A) Using ActionScript's `getURL()` within Flash content, and (B) using obfuscated JavaScript to redirect via `window.location`. Sophos products block these components as Mal/SWFRedir-A [24] and Troj/JSRedir-AZ [25] respectively.

It is normal for the URL of the SEO page to be included within the query string of the initial redirect URL, which enables the attackers to track the SEO pages that users are hitting. In most of the attacks analysed, there are multiple levels of redirection involved before the victim is exposed to the fake anti-virus site. A log of the HTTP traffic in an example attack is shown in Figure 9.

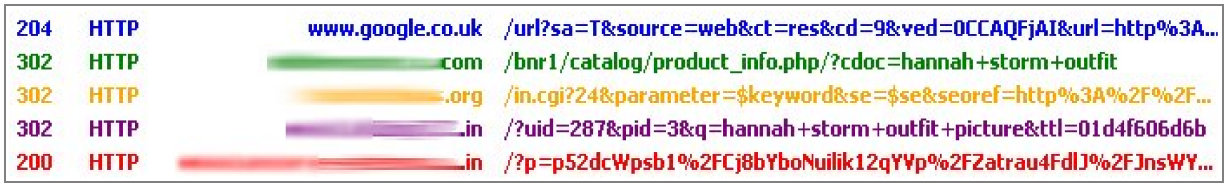

Figure 9: HTTP traffic observed after clicking through to an SEO page from Google search results. Multiple 302 redirects are used, from the SEO page (green) and two other rogue sites (orange, purple) before the user is taken to the fake anti-virus distribution site (red).

Thus far, it is clear that the attackers efforts to evade URL filtering technologies are focused on the second and third levels of redirection. The initial redirection (from the SEO page) is updated every 10 minutes for several of the kits analysed (via data retrieved from the C&C server), but only the sub-domain is changing. For the other kits the initial redirection has remained pretty static. We can predict this will likely change. Though we haven't seen it yet with SEO attacks, it would be trivial to embed an algorithm in the active content to alter the redirection (as seen with JS/Sinowal-Gen [26]).

4.5 Seeding SEO poisoned pages

Though the SEO poisoned pages may exist, they won't appear in search engine results until they have been indexed. In order for the search engine crawlers to index the page, they must first be aware that the page exists. This can be achieved by having other indexed pages link to it.

In most of the kits, each SEO page links to a number of other SEO pages so the process of seeding them for search engines to crawl is essentially automated (links to the SEO pages on different compromised sites is shared via the C&C). This means that only the first created SEO page in an attack needs to be seeded, which can can be conveniently done via forms provided by the search engines [27,28,29,30].

In kits where they choose not to link to other sites compromised in the same attack, we've observed a couple of other methods being used.

- The root page of the compromised sites is modified to contain links to the SEO pages hosted within that site. These links use simple CSS tricks to make them invisible to users, but to the search engine crawlers the links are readily indexed.
- Links are posted on other legitimate sites (through various user input mechanisms, such as comments, social bookmarking and blogging).

5 Protecting against SEO attacks

As for other web-based attacks, it is the combination of URL filtering and content inspection that provides the best protection for users against SEO attacks. Monitoring the currently active SEO attacks enables collection of the redirection URLs involved, which can then be appropriately blacklisted.

Detection-wise, protection can be provided by adding detection for the payload. As is typical for web hosted payloads, the binaries are frequently updated, and so a generic approach to detection is required. In addition to detection of the payload, analysis of each SEO attack can identify other components which may be detect-worthy, to thwart the attack prior to the user ever getting to the payload. Examples of such components include:

- active content used for redirection (as discussed in section 4.4)
- HTML elements/scripts used in malware distribution pages (e.g. Mal/FakeAvJs-A [31])

Providing detection for all the relevant components provides the most in-depth protection, with increased resilience against the server side polymorphism techniques that are used to obfuscate and change the script and binary content involved. This point is perhaps best illustrated by taking a look at some feedback data from *Sophos* customer web appliances. Table 3 summarises the top three detections across a 72 hour snapshot of data. Detections for the web pages used in fake anti-virus sites and the scripts used in SEO attack redirections are first and third most prevalent respectively.

Threat	Proportion of total threats detected
Mal/FakeAvJs	24%
Mal/Iframe	18%
SEO redirect scripts	9%

Table 3: Top three detections by prevalence across a 72 hour snapshot of data fed back from Sophos web appliances.

There is also the question of how site administrators protect their sites against attack, in order to avoid an SEO attack being hosted from within their site. As noted in section 3, based on the material collected during this investigation it is clear that vulnerable versions of popular CMS applications are a common link between many of the victim sites identified.

It is imperative that site administrators upgrade and patch such applications regularly. The homogeneous nature of the content produced by these CMS systems makes it trivial for attackers to identify potential sites to compromise. Once identified, there is no shortage of tools readily available to scan the site for vulnerabilities [32].

Content scanning on the web server can also add significant protection against SEO attacks, providing detection for the scripts used in SEO kits and PHP backdoors. Such detections can give administrators an early heads up of a potential server compromise.

6 Conclusion

In this paper we have provided insight into the current spate of SEO attacks by presenting the results of analysis into some of the kits being used to manage the attacks. Compromised legitimate websites provide a convenient network of hosts, that are being used as a platform for these attacks. By successfully poisoning search engine data, the attackers are able to lure unsuspecting users to malicious SEO pages, initiating the attack. To date, the attacks are being used for distribution of fake anti-virus malware, though the payloads could easily change in the future.

Similarly to how kits are used to automate and manage malicious content designed to exploit browser vulnerabilities [33], so too kits are being used to manage the SEO attacks. This facilitates setting up new attacks on different hosts, dynamically generating SEO pages stuffed with topical keywords and phrases for search engines to index. The kits can also provide a single point of control over the redirection URL used for SEO attacks hosted on multiple compromised sites.

Some of the kits provide functionality to automatically track the most popular search terms at any given time, in order to focus the attack on topical news and events, increasing the likelihood of attracting user traffic.

Malware distribution through SEO attacks could easily be described as beautiful in its simplicity. A straightforward case of trickery, without the need for exploits or zero-day vulnerabilities. Just a case of tricking the search engines into indexing rogue SEO pages and then tricking users into running the fake anti-virus malware (and subsequently paying to register it). This simplicity should not detract from the success of such attacks however. And whilst the attacks continue to succeed, there is little need for the malware authors and distributors to change the formula.

1 http://www.google.com/support/webmasters/bin/answer.py?hl=en-uk&answer=34444
2 http://help.live.com/help.aspx?mkt=en-GB&project=wl_webmasters
3 http://help.yahoo.com/l/us/yahoo/search/basics/basics-18.html
4 http://www.sophos.com/sophos/docs/eng/marketing_material/samosseiko-vb2009-paper.pdf
5 http://www.sophos.com/blogs/sophoslabs/v/post/844
6 http://www.sophos.com/support/knowledgebase/article/110379.html
7 http://www.sophos.com/support/knowledgebase/article/110379.html
8 http://www.sophos.com/blogs/sophoslabs/v/post/6765
9 http://www.sophos.com/blogs/sophoslabs/v/post/4308
10 http://www.sophos.com/blogs/sophoslabs/v/post/558
11 http://www.sophos.com/blogs/sophoslabs/v/post/250
12 http://www.sophos.com/blogs/sophoslabs/v/post/1329
13 http://www.sophos.com/blogs/sophoslabs/v/post/916
14 http://www.user-agents.org
15 https://addons.mozilla.org/en-US/firefox/tag/user%20agent
16 https://addons.mozilla.org/en-US/firefox/tag/referrer
17 http://en.wikipedia.org/wiki/Link_exchange
18 http://php.net/manual/en/book.curl.php
19 http://php.net/manual/en/function.fsockopen.php
20 http://www.google.com/trends/hottrends/atom/hourly
21 http://search.twitter.com/
22 http://blekko.com
23 http://www.w3.org/Protocols/rfc2616/rfc2616-sec14.html#sec14.30
24 http://www.sophos.com/security/analyses/viruses-and-spyware/malswfredira.html
25 http://www.sophos.com/security/analyses/viruses-and-spyware/trojjsrediraz.html
26 http://www.sophos.com/blogs/sophoslabs/v/post/8315
27 http://www.google.com/addurl/
28 http://www.bing.com/webmaster/SubmitSitePage.aspx
29 http://search.yahoo.com/info/submit.html
30 http://webmaster.yandex.ru/
31 http://www.sophos.com/security/analyses/viruses-and-spyware/malfakeavjsa.html
32 http://packetstormsecurity.org/
33 http://www.sophos.com/blogs/sophoslabs/post/3632