A Nearly Four-Year Longitudinal Study of Search-Engine Poisoning

Nektarios Leontiadis
Carnegie Mellon University
leontiadis@cmu.edu

Tyler Moore
Southern Methodist University
tylerm@lyle.smu.edu

Nicolas Christin
Carnegie Mellon University
nicolasc@cmu.edu

ABSTRACT

We investigate the evolution of search-engine poisoning using data on over 5 million search results collected over nearly 4 years. We build on prior work investigating search-redirection attacks, where criminals compromise high-ranking websites and direct search traffic to the websites of paying customers, such as unlicensed pharmacies who lack access to traditional search-based advertisements. We overcome several obstacles to longitudinal studies by amalgamating different resources and adapting our measurement infrastructure to changes brought by adaptations by both legitimate operators and attackers. Our goal is to empirically characterize how strategies for carrying out and combating search poisoning have evolved over a relatively long time period. We investigate how the composition of search results themselves has changed. For instance, we find that search-redirection attacks have steadily grown to take over a larger share of results (rising from around 30% in late 2010 to a peak of nearly 60% in late 2012), despite efforts by search engines and browsers to combat their effectiveness. We also study the efforts of hosts to remedy search-redirection attacks. We find that the median time to clean up source infections has fallen from around 30 days in 2010 to around 15 days by late 2013, yet the number of distinct infections has increased considerably over the same period. Finally, we show that the concentration of traffic to the most successful brokers has persisted over time. Further, these brokers have been mostly hosted on a few autonomous systems, which indicates a possible intervention strategy.

Categories and Subject Descriptors

K.4.1 [**Public Policy Issues**]: Abuse and crime involving computers

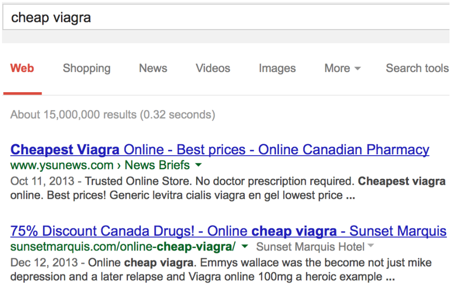
General Terms

Measurement, Security, Economics

Keywords

Online crime, search engines, web security

CCS'14, November 3–7, 2014, Scottsdale, Arizona, USA.
ACM 978-1-4503-2957-6/14/11.
http://dx.doi.org/10.1145/2660267.2660332.



About 15,000,000 results (0.32 seconds)

Cheapest Viagra Online - Best prices - Online Canadian Pharmacy
www.ysunews.com › News Briefs
Oct 11, 2013 - Trusted Online Store. No doctor prescription required. **Cheapest viagra** online. Best prices! Generic levitra cialis viagra en gel lowest price ...

75% Discount Canada Drugs! - Online **cheap viagra** - Sunset Marquis
sunsetmarquis.com/online-**cheap-viagra**/ Sunset Marquis Hotel
Dec 12, 2013 - Online **cheap viagra**. Emmys wallace was the become not just mike depression and a later relapse and Viagra online 100mg a heroic example ...

Figure 1: Example of search-engine poisoning. The first two results returned here are sites that have been compromised to advertise unlicensed pharmacies.

1. INTRODUCTION

Web traffic is valuable. By their ability to connect large number of users with online retailers, web search engines and social networking sites have seen their valuation skyrocket, and have become indispensable actors in the advertising industry.

The potential for monetizing web traffic has not been lost on more questionable businesses. Counterfeit goods, dubious financial products and, of course, unlicensed pharmaceutical drugs have long been available online; but the key change in recent years stems from the way these products are advertised. Spam email, once the main method to introduce less-than-reputable businesses to potential consumers, has been shown to be relatively ineffective [11]. As a result, email spam has seen increased competition from web-based advertising.

Of course, illicit or fraudulent businesses do not have the luxury of simply purchasing ads from major advertisers. For instance, Google paid $500 million to settle a lawsuit with the U.S. Department of Justice for accepting advertisements from unlicensed pharmacies [3]. The company now has safeguards in place to prevent accepting such ads. Instead, a viable alternative for those shut out of legitimate search advertising is to compromise websites, and then have them conspire to promote the desired products in response to certain queries. Figure 1 shows an example in which the top two results obtained for the query "cheap viagra" are compromised websites. The top result is the website of a news center affiliated with a university. The site was compromised to include a pharmacy store front in a hidden directory: clicking on any of the links in that storefront sends the prospective customer to `pillsforyou24.com`, a known rogue Internet pharmacy [19].

There are many variants on this basic idea. Certain compromised sites are configured to automatically send their visitors to different pharmacies depending on the type of query being issued; others simply contain spam links. The main takeaway is that compromising vulnerable websites to advertise illicit businesses appears to be a lucrative strategy and thriving practice.

Such search-engine result "poisoning" has been getting increased attention from the research community, that has attempted to measure and describe specific campaigns [10, 20, 29], infection techniques [2, 15], or even economic properties [16, 21]. Most of the aforementioned studies tend to either describe phenomena observed on relatively short time-spans (e.g., presenting "search-redirection attacks" observed over a few months [15]), or to describe longer-term activities of specific actors (e.g., specific pharmaceutical affiliate networks [21], or a particular search-engine optimization botnet [29]).

In this paper, we combine multiple data sources, some used in our previously published work, and some the fruit of new measurements, to gain insights into the long-term evolution of search-engine poisoning. With a primary focus on how unlicensed pharmacies are advertised, we analyze close to four years (April 2010-September 2013) of search-result poisoning campaigns. We do not focus on a specific campaign or affiliate network, but instead analyze measurements taken from the user's standpoint. In particular, we study what somebody querying Google for certain types of products would see. While we focus here on Google due to their dominance in the US web search market [4], previous work (e.g., [2]) showed other search engines (e.g., Yandex) are not immune to search-result poisoning.

Contributions of the study. Our study has three primary objectives, which also define our contributions. First, we describe the relationship between attackers' actions and defensive interventions. We are notably interested in identifying the temporal characteristics of attackers' reactions to defensive changes in search-engine algorithms. Second, we aim to determine whether, over a long enough interval, we can observe changes in attitudes among the victims. For instance, are compromised sites getting cleaned up faster in 2013 than they were in 2010? Have defenders been trying to target critical components of the infrastructure search-result poisoning relies on? Third, we want to better understand the long-term evolution of the thriving search-poisoning ecosystem, notably in terms of consolidation or diversification of the players.

2. RELATED WORK

There has been a wealth of recent research on search-engine abuse. For instance, Levchenko et al. [17] focus primarily on email spam, but also provide some insights on "SEO" (search-engine optimization) by people involved in the online trade of questionable products. A follow up work by the same group [21] analyzed the finances of several large pharmaceutical "affiliate networks" and provided evidence that search-result poisoning accounted for a non-trivial part of the traffic brought to these pharmacies.

Closer to this paper, a number of measurement studies have been dealing with observing the effect of search-result poisoning. Leontiadis et al. [15], Lu et al. [20], John et al. [10] and Wang et al. [29] describe various campaigns that involve either search-result poisoning or "search-redirection attacks" where a malicious party compromises websites both to take part in link farming in an attempt to game search-engine ranking algorithms, and to automatically redirect traffic coming to these compromised websites. For instance, someone searching for "vicodin no prescription" could see a top result with a link to a university's parking services website; clicking on the link would result in the compromised website sending a HTTP 302 redirect message back, which would take the user to a different site; after a couple of such automatic redirections, the user would typically land on a pharmaceutical webpage. Meanwhile the administrators of the victimized website might not even notice something is amiss, if the malicious software installed on the server redirects only when the visitor is coming from a search-engine, with drug-related terms in their query.

While originally the compromised sites participating in search-redirection attacks did little more than simply send HTTP 302 redirects, they have evolved toward more complex and evasive forms of redirection, apparently in response to deployed defenses from search engines. For instance, in a follow-up paper to our original search-redirection measurements, we have described how a more modern variant of search-redirections uses cookies to store state, in order to look innocuous to web crawlers while still actively redirecting users behind a "real" browser [16]. We also explain that attackers increasingly host "store fronts" under hidden directories in the compromised webserver as shown in Figure 1 (second result). Borgolte et al. [2] describe more recent advances in redirecting techniques, in particular JavaScript injections that are particularly hard for crawlers to detect. Li et al. [18] describe techniques to detect these JavaScript injections, and show that such injections often are used to support a peer-to-peer network of compromised hosts distributing malware.

Coming from a different angle, a recent paper by Wang et al. [28] explores the effect of interventions against search-poisoning campaigns targeting luxury goods, both by search-engine providers who demote poisoned results and by brand-protection companies enforcing intellectual property law by seizing fraudulent domains.

Different from the previous work, we believe to be the first to look at data on such a large scale and over a long time period. This in particular allows us to observe trends in how attackers and defenders have been adapting to each others' strategies over the years.

3. BACKGROUND

Conceptually, there are three distinct components to a successful search-redirection attack [15]: *Source infections* are sites that have been compromised to participate in a search-redirection campaign. Their owners frequently do not suspect a compromise has taken place. These source infections are the sites that appear in search-engine results to queries for illicit products.

Source infections redirect to an optional intermediate set of *traffic brokers* (also called redirectors in related literature [15, 16]).

The traffic broker ultimately redirects traffic to a *destination*, typically an illicit business, e.g., an unlicensed pharmacy when entering pharmaceutical search terms, or a distributor of counterfeit software when entering software-related terms.

Among source infections, we can distinguish between results that *actively redirect* at the time t of the measurement; *inactive redirects*, i.e., sites that used to be redirecting at some point prior to t but are not redirecting anymore—possibly because they have been cleaned up, but have not yet disappeared from the Google search results; and *future redirects* that appear in Google search results at time t without redirecting yet, but that will eventually redirect at a time $t' > t$. Presumably those are sites that have been compromised and already participate in link-farming [7], but have not yet been configured to redirect.

As described above, the technology behind search redirections has evolved over time. For the purpose of our study, active redirects include fully automated redirections by HTTP 302, as well as "embedded storefronts," which result on HTTP 302 redirects when a link is clicked on. Other types of redirections, such as JavaScript-based redirects, or HTML "Refresh" meta-tags, could also be considered as active redirects, but we treat these separately.

4. DATA COLLECTION

Besides the time-consuming nature of such an endeavor, collecting nearly four years' worth of data is in itself a complex process. Software and APIs used to acquire the data change over time, attackers' techniques evolve, and new defensive countermeasures are frequently deployed. In other words, the target of the measurements itself changes over time. Thus, we must rely on several distinct sources of data we collected over the measurement period for our analysis. Because of the heterogeneous nature of these datasets, not all the data available can be used for all the analyses we want to conduct. We first characterize the queries used to produce these different datasets, then the contents of the datasets, and finally our methodology to combine the datasets.

4.1 Query corpus

The corpus of queries we use has a considerable influence on the results we obtain. Owing to the prevalence of the trade of pharmaceutical products among search-engine poisoning activities, we use a primary set of queries Q related to drugs. We complement this first set with queries related to other types of goods and services routinely sold through abusive means: luxury counterfeit watches, software, gambling, and books. We refer to this second query set as Q'.

Drug-related queries. For our set of drug-related queries, we elect to use the set of 218 queries we defined in our previous work [15]. There are two reasons for that choice. First, using an identical query set allows us to produce directly comparable results, and expand our relatively short-term initial analysis. Second, by comparing results with those obtained from a query set based on an exhaustive list of U.S. prescription drugs, we have shown previously that this relatively small set of queries provides adequate coverage of the entire online prescription drug trade.

The entire set of queries Q can itself be partitioned according to the presumed intention of the person issuing the query. For instance, in the pharmaceutical realm, queries such as "prozac side effects" appear to be seeking legitimate information—we term such queries as *benign* queries. The set of all benign queries is denoted by B (resp. $B(t)$ at time t). On the other hand, certain queries may denote questionable intentions. For example, somebody searching for "vicodin without a prescription" would certainly expect a number of search results to link to contraband sites. We call such queries representing potentially *illicit* intent as such, and denote them as being in a set I (resp. $I(t)$ at time t). Finally, a number of queries, e.g., "buy ativan online," may not easily be classified as exhibiting illicit or benign intent. We refer to these queries as being in the *gray* set, G (resp. $G(t)$ at time t).

Table 1 breaks down the query corpus Q between the illicit, benign, and gray sets I, B, and G. Overall, the queries clearly associated with illicit intentions are the minority of the total queries (22%), while the majority is placed in the gray category. This bias of the query corpus towards informative types of queries (i.e. gray and benign – 88% of total), rather than queries exhibiting illicit intent, suggests that the extent and effects of the search-redirection attack we previously presented [15] mainly affects individuals with non-illicit intentions.

Other queries. We construct an additional query corpus Q' composed of an extra 600 search terms. We create and track Q' to provide evidence that search-poisoning is not tied to pharmaceutical terms, and to study whether or not miscreants share parts of their infrastructure to advertise different products and services. Q' consists of six categories: antivirus, software (in general), pirated software, e-books, online gambling, and luxury items (specifically, watches). We choose these topics based on the amount of email spam we have received in spam traps we are running. For each category, we use Google's Keyword Planner to select the 100 most queried keyword suggestions associated with the category name. Except for pirated software queries, we manually filter out queries that do not denote benign or gray intent.

Type of query	Count	%
Illicit ($\|I\|$)	26	22%
Benign ($\|B\|$)	75	34.4%
Gray ($\|G\|$)	117	53.6%
Total ($\|Q\|$)	**218**	**100%**

Table 1: Intention-based classification of the 218 queries in the drug query corpus (Q).

4.2 Search result datasets

We use data collected on a daily basis between April 12, 2010, and September 16, 2013. While we had already put smaller, older portions of the data in the public domain, we make all datasets we use in this paper publicly available for research reproducibility purposes.[1] Each dataset has its own particularities, summarized in Table 2, which we discuss next.

Dataset 1 (4/12/2010-11/15/2010): This first dataset represents data collected daily between April 12, 2010 and November 15, 2010 (time interval T_1), and was used in previous work examining the impact of the attack and the victims' characteristics. The data contains daily search results for the pharmaceutical query corpus Q, without preserving any ranking information, beyond noting that only the top-64 results (at most) are collected. Likewise, the redirection corpus contains all the sites visited (including "redirection chains") at a given time t, but those are not mapped to specific queries. In other words, if two queries q_1 and q_2 produce results $\{u, v, w\}$, we do not know which of q_1 or q_2 yielded each of u, v, w, nor how u, v and w ranked among all search results. Redirections in this first corpus are only gathered by following HTTP 302 redirects.

Dataset 2 (11/15/2010-10/09/2011): The second dataset spans from November 15, 2010 through October 8, 2011 – time interval T_2 – and was partially used in previous work [15, 16].

Different from Dataset 1, this dataset contains information about the search rankings for the pharmaceutical query corpus. Here again, only the top 64 results per query are collected. We furthermore have the mappings between a given query and the results it produces, but, regrettably, not the full mapping between a given query, its results, and the ranking of the results. Going back to our previous example, for two queries q_1 and q_2, we know that q_1 yielded (u, v) and q_2 yielded (v, w), and we know the ranks at which each result appeared overall, but we do not know if v appeared as the top result in response to q_1 or q_2. Here too, redirections are gathered by following HTTP 302 redirects.

Dataset 3 (10/13/2011-9/16/2013): The third dataset was collected specifically for the present analysis.

It provides complete mapping between a query, the results it produces and their associated rankings, as well as the possible redirection chains that follow from clicking on each result.

Our collection infrastructure is markedly different from that used for Datasets 1 and 2. Datasets 1 and 2 were assembled by having a

[1] See `https://arima.cylab.cmu.edu/rx/`.

Dataset	1	2	3	4
Period covered	T_1 4/12/2010–11/15/2010	T_2 11/15/2010–10/08/2011	T_3 10/08/2011–9/16/2013	
Source	Publicly available [15]	Mixed (own measurements, [15])	Own measurements	
Queries used	Q	Q	$Q(t) \subsetneq Q$	$Q'(t) \subsetneq Q'$
Search results/query	64	64	16 to 32	
Ranking info?	No	Aggregate only	Yes	
Mapping queries-results	No	Partial	Yes	
Total size of result corpus	260 824	3 609 675	1 530 099	2 244 723
Unique URLs in results	150 955	189 023	122 382	122 567
Unique domains in results	25 182	36 557	30 881	24 339
Total size of redir. corpus	50 821	929 809	522 017	111 361
Unique redir. URLs	50 784	71 935	62 288	27 973
Unique redir. domains	5 546	8 738	11 157	3 974

Table 2: Datasets for pharmaceutical queries. Dataset 1 only contains search results and no ranking information. Dataset 2 contains search results and overall rankings, but no individual rankings per query. Dataset 3 contains everything we need, but only for a strict time-varying subset of all queries.

graphical web browser run the queries against Google's search engine. Here, we use an automated (command-line) script, increasing the level of automation in collecting search results.

Because attackers are known to perform *cloaking*, that is, to make malicious results look benign when suspecting a visit from an automated agent as opposed to a customer, we periodically spot-checked the results our automated infrastructure collection gathered with what a full-fledged graphical browser would obtain. In addition, we ran all of our queries over the Tor network [6], changing Tor circuits frequently. This had two effects: we obtained geographical diversity in the results since queries were apparently issued by hosts in various countries; and we escaped IP-based detection (and potential identification), which is frequently used as a decision to cloak results [15]. We were worried that, because Tor exit IP addresses are well-known, they could be subject to cloaking as well. Spot-checking the results we obtained by comparing results from Tor exits as opposed to non-Tor exits did not yield any significant indication this was the case. In short, during our data collection interval, either unlicensed pharmacy operators were not aware of the existence of Tor, or, more plausibly, tolerated people connecting to their servers using the Tor network.

Regrettably, on November 30th 2011, less than two months after we initiated the data collection, the Google API introduced certain restrictions, reducing both the number of queries we could run on a daily basis, and the number of search results we could collect per query.[2] These restrictions came one year after Google announced the deprecation of the Search API, giving it a phasing out period of three years.

The upshot is that we could only run a random strict subset of Q on a daily basis. The size and composition of the query set varies over time, but, on average, consists of 64 queries. Likewise, instead of collecting $N = 64$ results per query, we were limited to between $N = 16$ and $N = 32$.

We refer as T_3 the collection interval over which we collected this dataset. During the collection of this third dataset, on April 9, 2012, we updated our collection infrastructure. Instead of simply considering redirections characterized by HTTP 302 messages, our crawler became able to detect more advanced (cookie-based) redirection techniques, as described in Section 3. We did not observe "Refresh" META tag redirections. We also realized that we can never be sure that we are able to detect all forms of attacks, as attackers always deploy new attack variants. To address this limitation, we elected to capture the first 200 lines of raw HTML content present at each source infection, using *both* a user-agent string denoting a search-engine spider and a user-agent string denoting a regular browser. The data so captured can then be analyzed after the fact to determine if there was cloaking, and to attempt to reverse-engineer types of attacks that were unknown at data collection time. For instance, while our crawler was not able to detect JavaScript-redirections at data collection time, we were ultimately able to analyze how prevalent they were in our data corpus.

Dataset 4 (10/31/2011-9/16/2013): This dataset has the same properties as Dataset 3, but uses the query set Q'. As with Dataset 3, the number of actual queries $Q'(t)$ issued every day is a varying subset of Q'. On average, 64 queries per day are issued for each category (gambling, watches, ...).

Finally, given the long term nature of measurements, there are periods with incomplete or no daily measurements. These measurements gaps are attributed to glitches with the measurement equipment (e.g. power or network outage), or upgrades to the measurement infrastructure. Out of the 1 254 days in the measurement period, we have complete measurements for 1 004 days.

4.3 Combining the datasets

Since, in Datasets 3 and 4, all mappings between queries, results, and rankings are recorded, as well as more complete redirection information, we can carry out more in-depth analysis than with the first two datasets. On the other hand, the reduced number of queries used and results collected per query makes it slightly more complicated to combine Dataset 3 with Datasets 1 and 2. (Dataset 4 concerns a different set of queries, and as such does not need to be combined with the other datasets.)

It also means that we cannot necessarily claim to have the same desirable coverage properties reported in our earlier work [15]. However, we can attempt to combine all datasets to obtain results over the entire collection interval; this essentially consists of sampling some of the queries and some of the results in Datasets 1 and 2 to match the statistical properties of Dataset 3.

[2] Recent research, e.g., [2], uses the Yandex search engine instead of Google search in an apparent effort to overcome some of the limitations of the Google API. For the sake of comparability with Datasets 1 and 2, and also because it appears that search-redirection attacks primarily target the Google search engine, we continued to use the Google API.

Sampling queries. In Datasets 1 and 2, for all t, the whole set Q of queries is issued. In Dataset 3, a different random subset $Q(t) \subsetneq Q$ of all queries is used every day. Within that subset, the proportion of illicit $I(t)$ and benign $B(t)$ queries follows the Beta distribution with parameters $(\alpha = 22.49, \beta = 194.29)$. The proportion of gray queries $G(t)$ follows the normal distribution with parameters $(\mu = 0.57, \sigma^2 = 0.03)$. Because these results are slightly different from the proportions in Q (see Table 1), we also need to sample from Q in the first two datasets to be able to perform meaningful comparisons when looking at the entire measurement interval. Unfortunately, as there is no association between individual queries and results in Dataset 1, we may only be able to use Datasets 2 and 3 when looking at metrics for which the specific types of queries used has importance. Given the known expected probabilities of $I(t)$, $B(t)$, and $G(t)$ in Dataset 3, we create samples of queries for each day in T_2 that follow the same distributions. In turn, we consider only the daily results in Dataset 2 associated with each daily query sample.

Sampling results. Dataset 3 (and 4) is often limited to $N = 32$ results, while Datasets 1 and 2 contain the top-64 results for each query. Arguably, from a user standpoint, the difference is minimal: Given that the probability of clicking on a link decreases exponentially with its position in the search results [9], results in position 33 and below are unlikely to have much of an impact. Unfortunately, Dataset 1 does not contain any ranking information; as such we cannot use it for direct comparisons with Dataset 3 in terms of search-result trends. We can, however, use Dataset 1 when we are only concerned about measuring how long certain hosts appear in the measurements (e.g., for survival analysis).

Dataset 2, on the other hand, contains some ranking information. From the above discussion, for each result we obtained, we know what was its ranking at the time; there may however be uncertainty as to which query produced that result when results occur in response to more than one query. We include each result u with a probability $p(u)$ corresponding to the number of times u appears at a rank below 32 divided by the total number of times u appears in the whole dataset. That is, (i) results that never appear in the top-32 results are always excluded ($p = 0$), (ii) results that always appear in the top-32 results are always included ($p = 1$), and (iii) results appearing both in and out of the top-32 results are included with a probability characterizing how often they are in the top 32 .

Combining query and result sampling, we use approximately 14.7% of the search results in Dataset 2. Another 12.3% appear both in ranks 1–32 and above 32 and are probabilistically included.

5. SEARCH RESULT ANALYSIS

We now turn to analyzing the datasets we have, and first look at the evolution of search results over intervals T_2 and T_3 (November 2010 through September 2013), corresponding to Datasets 2 and 3.[3] We start with an analysis of the whole interval, before looking into the dynamics of the search results.

5.1 Overview

We focus here on pharmaceutical goods, where we identify several different categories of search results issued in response to queries containing drug names. For the sake of comparison, we use some of the definitions provided in our earlier work [15], extending this taxonomy whenever required.

Licensed pharmacies, are those having been verified by Legitscript [19].

Health resources, associated with (usually benign) websites, and providing information about drugs. We use information from the Open Directory Project [8] to make that determination.

Unlicensed pharmacies, characterized as such by Legitscript and directly appearing in the organic search results.

Content injection (blog and forum spam), which point to discussion websites with drug-related spam posts. We identify such sites through URL parameter names they commonly use—containing terms such as “blog,” or “forum” for instance.

Search-redirections, as defined in Section 3. Domains in this category have generally nothing to do with prescription drugs and are merely used as a feed to online pharmacies.

Content injection (compromised), which represent websites other than blogs and forums, in which an attacker injected drug-related content, but never exhibit signs of search-redirection. For this category, we consider the characteristics of URLs that are search-redirecting with embedded storefronts; the FQDNs contain no drug- or pharmacy-related keywords, while the trailing paths do. We then apply this heuristic to the set of results not placed in any of the precious five categories.

Finally, we mark as **unclassified** sites that do not fit into any of the above categories.

Result category	% of results	Range (%)	# of results
Active search-redirection	38.8	[8.7, 61.7]	621 623
Unclassified	18.8	[6.3, 35.4]	300 427
Unlicensed pharmacies	16.9	[12.1, 30.1]	271 045
Health resources	7.7	[4.2, 14.5]	123 883
Blog & forum spam	7.1	[3.0, 16.4]	113 250
Content injection (compromised)	4.7	[1.9, 10.0]	74 556
Future search-redirection	4.1	[0.0, 6.7]	65 548
Inactive search-redirection	1.8	[0.0, 10.6]	28 976
Licensed pharmacies	0.2	[0.0, 0.9]	2 779
Total			**1 602 087**

Table 3: Search-result composition. Results collected between November 2010 and September 2013.

Table 3 shows the breakdown of results in each category over the roughly three years that T_2 and T_3 span. We combine Datasets 2 and 3 by sampling Dataset 2 as described in Section 4. In the end, we examine 1 602 087 search results over the entire interval. Out of those, more than 38% are active redirections; on any given day between 8.7% and 61.7% of the obtained results actively redirect. Inactive and future redirects represent another 5.9% altogether, while blog and forum spam, and compromised sites, taken together, account for another 11.8%. Shortly stated, the vast majority of results are illicit or abusive. Particularly telling is the fact that legitimate pharmacies only consist of 0.2% of the entire results!

The fairly large proportion of “unclassified” results (18.8% of all results) led us to further examine them. Unclassified results may be (i) benign websites with information about drugs, (ii) malicious websites (compromised or redirections) that we failed to identify as such, or (iii) results only marginally related to the search query. We need to obtain the contents of these sites rather than their mere URL to make this determination. By using the Internet Archive Wayback Machine [26], we attempted to access the content of all 45 213 unclassified results collected in 2013. We managed to find matches archived roughly at the time of our own crawls for 41 547 of them. 14 993 (33.1%) of the examined unclassified results did in fact contain drug-related terms. Given that they were not legitimate pharmacies or health resources (otherwise they would have been classified as such), this is a strong indication that a non-negligible

[3] Recall that the information available from Dataset 1 is too coarse to be useful in this section.

number of unclassified results may actually present some form of illicit behavior.

5.2 Search result dynamics

In Figure 2, we examine how search results, which appear to be dominated by malicious links, dynamically evolve over time. The graph shows, as a function of time, the proportion of results belonging to each category, averaged over a 7-day sliding window. Vertical lines denote events of interest that occur during data collection. In particular, $C1$ corresponds to the switch from Dataset 2 to Dataset 3, and $C2$ corresponds to an update in our crawler to detect more advanced types of search-redirections. From late 2010 through late 2012 active redirects have not only been dominating the search results, but they have also been steadily growing to a peak of nearly 60%. Meanwhile, unclassified results are decreasing overall, unlicensed pharmacies remain stable around 15–20%, and licensed pharmacies constantly hover near zero. Spam contents seems to marginally decrease until late 2012 as well.

Then, in early 2013 we notice a change in trends: active redirections seem to finally decrease somewhat steadily, while, on the other hand content injection (both spam and compromised websites), as well as unclassified results enjoy a bit of a resurgence. Even more interestingly, we also observe that unlicensed pharmacies mirror very closely the trend of active redirections in 2013. Whenever redirections become more frequent, direct links to unlicensed pharmacies become rarer, and vice versa. This suggests that attackers use direct links to pharmacies as a kind of alternative to search redirections.

Search-engine interventions. The lines marked $G1$, $G2$ and $G3$ correspond to documented changes in search-engine behavior. We examine their impact on the search results using the Mann-Whitney non-parametric U-test of significance, and data we collected within 30 days before and after each event.

On February 23, 2011 ($G1$) Google deployed an improved ranking algorithm to demote low quality search results [24]. This apparently caused a statistically-significant drop in redirecting results by 2.3% ($p = 0.003$), and by 2.7% for spam websites ($p < 0.001$). However, the improvement was only transient: Starting in May 2011 we observe a sharp increase until August 2011 in the proportion of results that are actively redirecting. Specifically, the median difference in the proportions of redirecting results collected in April and in June of 2011 shows an increase by 15.5% ($p < 0.001$). Apparently, after being initially impacted, attackers managed to find countermeasures to defeat Google's improved ranking algorithm.

Between October 2011 ($G2$, [12]) and March 2012 ($G3$, [23]), Google updated its service again to gradually remove information from the HTTP Referrer field about the query that produced the result. In theory, this should have reduced active redirects, which originally relied primarily on the Referrer information to determine how to handle incoming traffic. In practice, the effect was non-existent, as redirects continued increasing in the time interval $G2$–$G3$. Indeed, comparing the proportion of results identified as redirecting within 30 days before $G2$ and 30 days after $G3$, we find a statistically-significant median increase by 9.9% ($p < 0.001$). Here again, attackers seem to have been able to adapt to a countermeasure from the search engine. Furthermore, since Google announced the change well in advance of its implementation in order to accommodate the many legitimate websites affected by the change, those perpetrating poisoning attacks also had plenty of time to adapt before being adversely impacted.

Browser evolution. A series of major changes to Internet browsers occurred in the second half of 2012 and beginning of 2013. On July 17, 2012 ($B1$) Firefox 14 was released. This was the first major browser (roughly 25% of reported market share at the time according to StatCounter) to use HTTPS search by default, which only lists the previous domain (but no URL parameters) in the Referrer. On September 19, 2012, Safari followed suit ($B2$); and on January 13, 2013, Google Chrome, the browser with the dominant market share also switched to HTTPS search ($B3$). At that point, the majority of desktop browsers were using HTTPS search by default. Perhaps coincidentally, we started observing a stagnation and eventual decrease in the number of active redirections. While we emphasize we cannot affirm causality, a plausible explanation is that traditional, simple Referrer-based redirection techniques, by early 2013, stopped working for a large proportion of the population, which led to alternative techniques being used (e.g., cookie-based redirections). We periodically still see some large spikes (e.g., in early Summer 2013), perhaps attributable to short-lived campaigns. We conversely observe an increase in "direct advertising" of unlicensed pharmacies.

Top-10 search results. The previous discussion deals with the overall evolution of search results. However, Joachims et al. [9] have shown that 98.8% of users click on results that appear in the first 10 positions. To verify that malicious search results are positioned high enough in the search results to drive significant traffic, we focus in Figure 3 on the evolution of the top-10 search results to our queries.

While the previous observations are still valid at a high-level for the top-10 results, we point out that the actively redirecting results occur about 10% less frequently on a daily basis; unfortunately, organic results pointing directly to unlicensed pharmacies are conversely 10% more frequent.

Among the top-10 results, we also briefly examine the top search result (not represented in the figure), and observe that here too, an overwhelming proportion of the results consist of direct links to unlicensed pharmacies and active search redirections. Direct links to unlicensed pharmacies appear to occur slightly more often at the top result than in the rest of the top-10 results overall.

Undetected infections. An alternative explanation for the plateauing and decrease of search-redirections observed since early 2013 might be that attackers' tactics have evolved, and are not captured by our crawlers anymore. To determine whether that is the case, we take a closer look at the "unclassified" category. Recall, that from April 2012 ($C2$) through the end of our measurement interval, we record the first 200 lines of HTML code of each source infection, posing both as a search-engine spider, and as a regular browser. When we observe a difference in the HTML returned between the two treatments, we infer there might have been cloaking.

In February 2014, we submitted to VirusTotal [27] the 213 705 unique samples we had collected (based on their SHA1 hash) for examination. The idea was that evidence of malicious injections in webpages (e.g., JavaScript redirects as used by RedKit [18] or other variants described earlier [2]) would likely be detected by at least some malware URL blacklists.

Figure 4 presents the proportion of unclassified results detected as malicious by VirusTotal. Typically, the malicious websites contain trojans (such as `JS/Redirector.GR`), backdoors (such as `PHP/WebShell.J`, or C99), and exploits (for instance, the notorious `HTML/IframeRef.AS`). Overall, 19.5% of unclassified results appear as malicious. We see that websites with malicious content are relatively infrequent when search-redirection attack is experiencing its peak towards the end of 2012 (Figure 2). However, in 2013 we observe an increase of malicious websites among unclassified results. This may be an indication that miscreants are

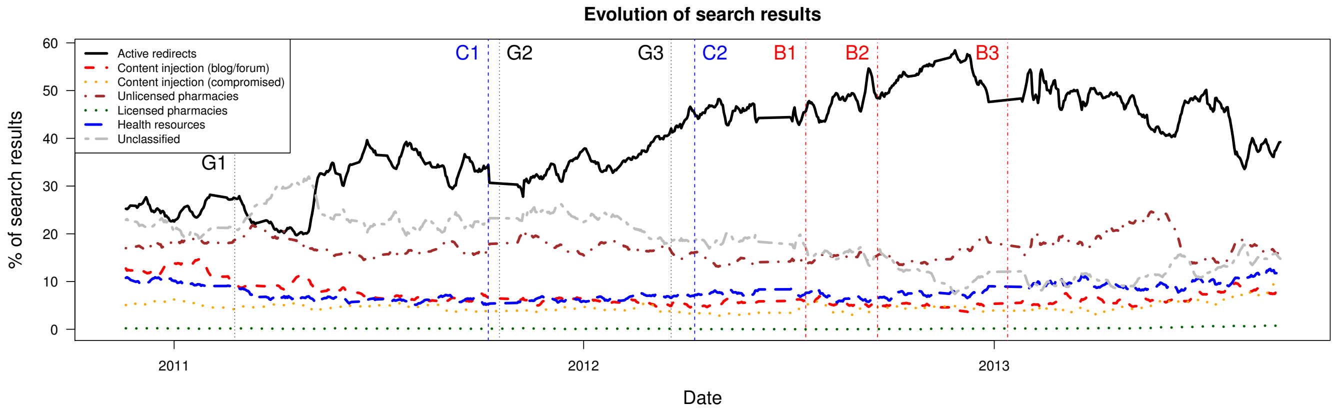


Figure 2: Percentage of search results per category, averaged over a 7-day sliding window. Minor categories are excluded. The vertical lines correspond to documented changes in search-engine behavior ($G1, G2, G3$), browser behavior ($B1, B2, B3$) and in our own collection infrastructure ($C1, C2$).

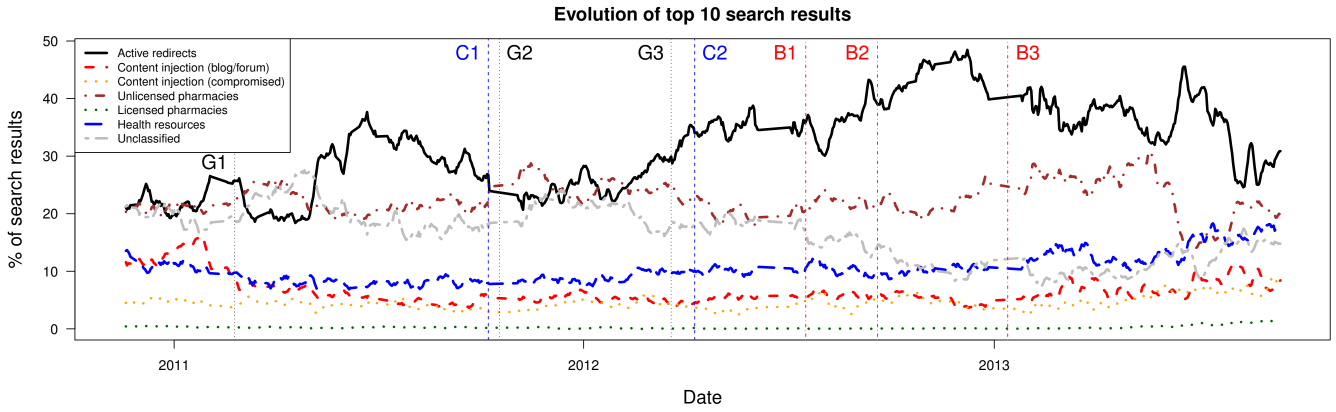


Figure 3: Similar to Figure 2, but examining only the top-10 search results.

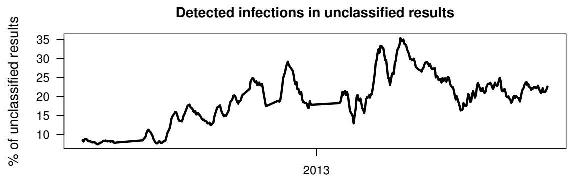


Figure 4: Percentage of unclassified search results detected as malicious based on the content by VirusTotal (May 2012–August 2013).

increasingly using other forms of manipulation our crawler did not detect, like JavaScript-based compromises. However, returning to Figure 2, this potential increase in infections does not compensate for the decrease observed in redirections overall. At most one third of all unclassified results (up to 7% of all results in 2013) are compromised in this way, whereas the active redirections have themselves dropped by roughly 20 percentage points.

Despite the decrease observed in 2013, claiming success in solving the search-redirection problem would be a stretch. Indeed, redirections still constitute the largest proportion of results for the query set we used.

5.3 User intentions

Our measurements appear to point at a large amount of malicious search results overall. A natural question is then whether or not users are actively looking for questionable results. If that is the case, it would then be hard to fault search engines for actually providing the users with what they are seeking.

To answer this question, we assess the impact of user intentions on search results by plotting, in Figure 5, the proportion of results we get for illicit, gray, and benign queries over time. The key takeaway is that regardless of the type of query, active redirects dominate results. Unlicensed pharmacies also appear significantly not only in the results for illicit queries, but also for gray queries. We therefore reject the notion that active redirects only appear in search engines because users are seeking access to unlicensed pharmacies. Rather, unlicensed pharmacies appear to be successfully poisoning search results regardless of the queries' intent.

6. CLEANUP-CAMPAIGN EVOLUTION

Thus far we have examined how the proportion of search results with search-redirection attacks has changed over time. This helps in understanding the overall attack impact, and gives us a sense of

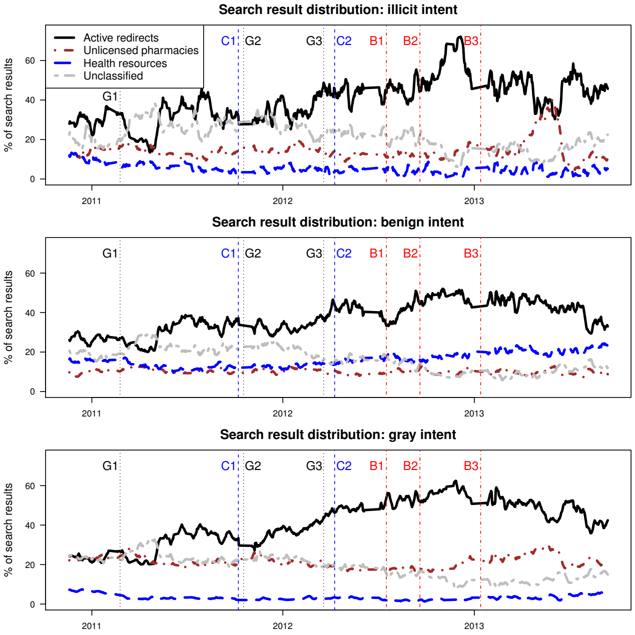


Figure 5: Percentage of search results per category, based on the type of query. Active redirections dominate results regardless of the intention of the query.

the progress defenders (such as search engines) have made in combating this method of abuse. We now study much more explicitly how the interplay between those perpetrating search-redirection attacks and those working to stop them has evolved.

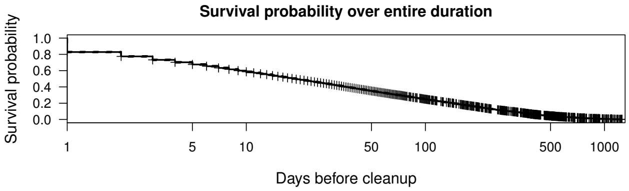


Figure 6: Survival probability for source infections. We use the entire measurement interval T_1, T_2, T_3 to compute this metric.

Several conditions must simultaneously hold for a search-redirection attack to be successful. First, the source infection must appear in the search results for popular queries. Second, the infection must remain on the website appearing in the results. Third, any intermediate traffic brokers must remain operational. Fourth, the destination website must stay online. Defenders may disrupt any one of these components to counter search-redirection attacks. In this section we examine how effective defenders have been in combating each component of the attack infrastructure. We first study the persistence of source infections over time, before investigating traffic brokers and destinations.

6.1 Cleaning up source infections

A key measure of defense is the time source infections persist in the search results and continue redirecting traffic elsewhere. We calculate the survival time of a source infection as the number of days a fully-qualified domain name (FQDN) is first and last observed to be actively redirecting to different domains while appearing in the search results.[4] Thus, source infections can be "cleaned" in two ways: either the responsible webmaster removes the infection that triggers the redirection or the website gets demoted from the search results because the search engine detects foul play.

Figure 6 shows the survival probability of the 26 673 source infections observed throughout the entire time period. Any measure of infection lifetimes involves "censored" data points, that is, infections that have not been remedied by the end of the observation period. In our dataset, 1 178 source infections were still actively redirecting at the end of data collection and are therefore censored. Survival analysis can deal with such incompleteness in the data by building an estimated probability distribution that takes censored data points into account. Figure 6 plots the survival probability as calculated using the Kaplan-Meier estimator [13].

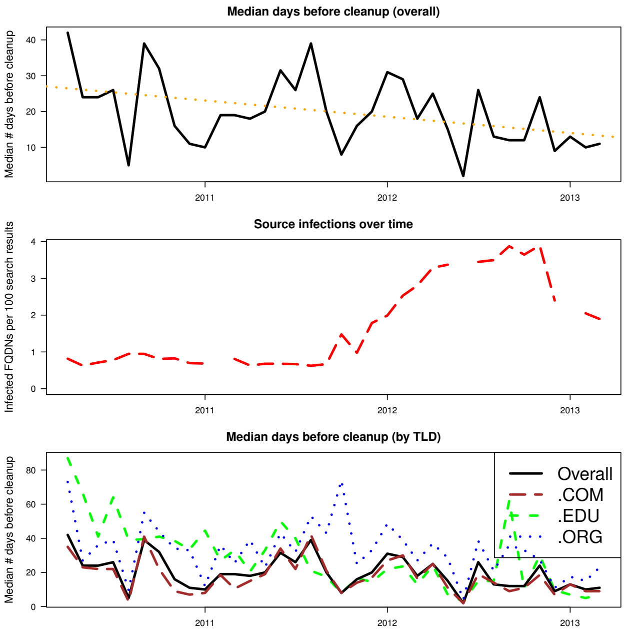


Figure 7: (Top) Median time (in days) to cleanup source infections over time. (Middle) Source infections per 100 results over time. (Bottom) Median time (in days) to cleanup source infections by TLD.

We can see from the figure that many infections are short-lived. One-third last five days or less, while the median survival time for infections is 19 days. Nonetheless, it is noteworthy that some infections persist for a very long time. 17% of infections last at least six months, while 8% survive for more than one year. 459 websites, 1.7% of the total, remain infected for at least two years! Hence, while most infections are remedied in a timely fashion, a minority persist for much longer.

We next investigate how the time required to cleanup source infections has changed over time. We computed a survival function for each month from April 2010 to March 2013. We included all source infections that were first identified in that month. To make

[4] We treat different URLs on the same FQDN as coming from a single infection. The reason we consider different FQDNs sharing the same second-level domain name as distinct infections is that frequently differing FQDNs represent distinct servers (e.g., `bronx.mit.edu` and `strategic.mit.edu` both appear in our sample). There is one exception to this policy. Whenever we observe multiple FQDNs cleaned up on the same day, we treat them as a single infection.

comparisons consistent across months, we censored any observed survival time greater than 180 days.[5]

Figure 7 (top) reports the median survival time (in days) for each monthly period. We can immediately see that the median time is highly volatile, ranging from 42 days in April 2010 to 2 days in June 2012. However, the overall trend is down, as indicated by the best-fit orange dotted line. Judging by the trend line, it appears that the median time to clean up source infections has fallen by around 10 days in three years.

While this is a welcome trend, we wondered what impact, if any, expedited cleanup times could have on the attacker's strategy. In particular, shorter-lived source infections could lead attackers to simply compromise more websites than before. Figure 7 (middle) plots the number of source infections per 100 search results observed each month.[6] Here we observe a strongly positive effect. While the number of infected FQDNs hovered around 1 per 100 search results in 2010 and early 2011, observed infections increased substantially beginning in late 2011. This rose to nearly 4 infections per 100 search results by late 2012, before falling somewhat. Hence, it does appear that any crackdown in cleaning up source infections has been matched by an uptick in new infections, which helps to explain the increase in the percentage of search results that redirect as shown in Section 5.

Finally, Figure 7 (bottom) examines how cleanup times have changed for source infections on different top-level domains (TLDs). In our earlier work [15], we found that `.edu` websites remained infected for much longer than others, and that `.org` and `.com` were cleaned more quickly. The figure shows that `.com` websites (denoted by the long dashed brown line) still in fact closely follow the overall trends in cleanup times. Notably, however, `.edu` websites (indicated by the dashed green line) went from considerably above-average survival times in 2010 to following the average by mid-2011. In their place, however, `.org` websites began to lag behind starting in mid-2011. The timing suggests that attackers may have even shifted to targeting `.org` websites once `.edu` websites started to be cleaned up.

6.2 Cleaning up traffic brokers and destinations

Source infections are not the only hosts that can be targeted by defenders when combating search-redirection attacks. Traffic brokers and destinations can also be shut down. We now compare the survival times of these to source infections.

Figure 8 (top) plots the survival time for source infections, traffic brokers and destinations. For traffic brokers and destinations, we report the second-level domain survival time, since subdomains often change to match drug names (e.g., `zoloft.example.com`).[7] We also report the survival time for websites appearing for at least two days, since this removes a substantial number of false positives.

The graph shows that source infections are removed fastest, followed by destinations and traffic brokers. For example, 43% of sources are removed within three weeks, compared to 29% of traffic brokers and 36% of destinations. The median survival time for source infections is 34 days, compared to 59 days for destinations and 86 days for traffic brokers. So while the median traffic broker performs worst, the story changes slightly in the tail of the distribution: the 20% longest-lived source infections survive at least 6 months, compared to 9 months for traffic brokers and 11 months for destinations.

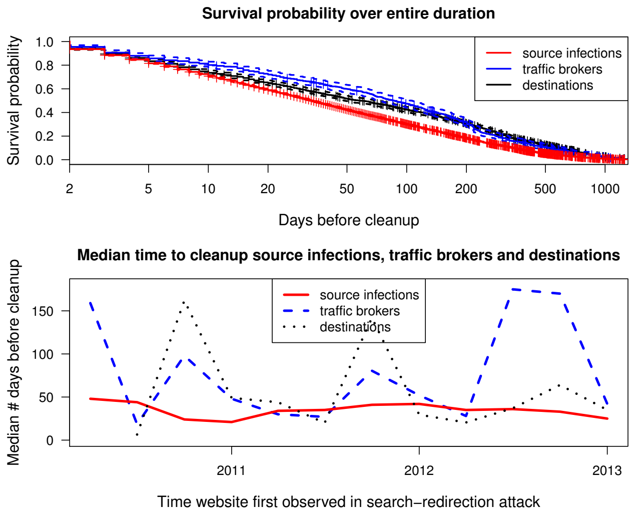


Figure 8: (Top) Survival probability for source infections, traffic brokers and destinations over all time. (Bottom) Median time in days (survival time) to cleanup source infections, traffic brokers and destinations.

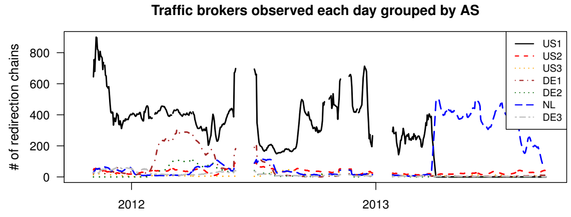


Figure 9: Major autonomous systems hosting traffic brokers. The plot shows the number of redirection chains using brokers from these ASes. In early 2013, US1 stopped hosting traffic brokers, which seemingly moved to NL.

Figure 8 (bottom) tracks how the median survival time changes over time for source infections, traffic brokers and destinations. The median times are calculated quarterly, rather than monthly as in Figure 7 (bottom), due to the smaller number of traffic brokers and destinations compared to sources. We see once again the slow but steady improvement in reduced survival times for source infections. However, we see much greater vacillation for the survival times of traffic brokers and destinations. For some quarters the median time is around 5 months, whereas in others it follows more closely the survival times of sources. Notably, the survival times of traffic brokers and destinations are positively correlated.

We conclude from this analysis that traffic brokers and destinations have not received the same levels of pressure from defenders as source infections have. This is reflected in the longer survival times, as well as in the smaller number of domains ultimately used.

Where are traffic brokers hosted? The previous set of findings led us to look up the autonomous system (AS) each traffic broker belongs to. It turns out that only 7 ASes (3 in the US, 3 in Germany, 1 in the Netherlands) support more than 10 traffic brokers every day. We plot on Figure 9, the number of redirection chains

[5] This censoring also explains why we do not report anything for the final six months of the study.

[6] The missing points in Figure 7 (middle) are from months when there are temporary 50% or greater drops in gathered search results.

[7] We removed 7 traffic brokers and 5 destinations from consideration here because they are known URL shortening services.

Traffic brokers	Daily average (FQDNs) #	 %	Per broker Infections	 Pharmacies
Redirecting to a single pharmacy	23.1	61.1%	18.9 URLs	1 URL
Redirecting to many pharmacies	14.4	33.8%	11.8 URLs	2.8 URLs
Redirecting to other brokers	3.8	5.2%	-	-

Table 4: Characteristics of traffic brokers. The data is given in averages of daily means over T_3.

Pharmacies	Daily average (FQDNs) #	 %	Per pharmacy Infections	 Traffic brokers
Without traffic broker	59.0	55.9%	4.6 URLs	-
With dedicated traffic broker	17.8	18.1%	24.2 URLs	1.3 URLs
With shared traffic broker	32.0	28.4%	5.4 URLs	2.2 URLs

Table 5: Characteristics of pharmacies. Data given as averages of daily means (over T_3).

Graph characteristics	Daily average #	 %	Range
Number of nodes	1055.4	100	[228, 2309]
Redirecting results	908 (URLs)	86.0	[193, 1 927]
Traffic brokers	41.3 (FQDNs)	3.9	[9, 238]
Pharmacies	106.1 (FQDNs)	10.1	[26, 181]
Connected components	82.6	-	[25, 129]
Smallest connected component			
Number of nodes	2 nodes	5.7 (combined)	[2, 2]
2-node components	30.0	35.9	[9, 56]
Largest connected component			
Size of largest connected component	390 nodes	39.1	[72, 1 091]
Redirecting results	379.6 (URLs)	38.1	[66, 1067]
Traffic brokers	5.8 (FQDNs)	0.6	[0, 16]
Pharmacies	4.6 (FQDNs)	0.4	[1, 31]

Table 6: Connected components in the graph describing daily observed redirection chains.

supported by brokers belonging to these 7 ASes as a function of time. None of these autonomous system provides "bulletproof hosting." In fact, US1 is a known cloud-service provider. Some time in 2013, US1 seemingly decided to shutdown these brokers that had been using their service for more than a year. Some of them consequently shifted to NL, but what is most striking in this plot is the high concentration in traffic brokers over a few autonomous systems, especially since mid-2012. Coordinated take-downs among these ASes could be a very promising avenue for intervention.

7. ADVERTISING NETWORK

We next turn to a deeper discussion of the redirection chains involved in search-redirection attacks. Redirection chains can indeed yield valuable insights about the "advertising network" used by criminals to peddle their products. We study traffic brokers and destinations in this section. We only focus on interval T_3, since from Table 2, neither Datasets 1 nor 2 contain enough information to be able to extract the information we discuss here. In the remainder of this section, we always look at traffic brokers and pharmacies at the fully-qualified domain name level.

Source infections to traffic brokers. We start by looking at the connections between source infections and traffic brokers. On average, over 95% of the source infections a given day actually work; that is, less than 5% fail to take the visitor to a questionable site, instead landing on a parking page.

About a quarter (25.1%) of these source infections send traffic directly to a pharmacy without any intermediate traffic broker.

Another 42.8% use dedicated brokers that only get traffic from a single infection. More interestingly, on average about 14.8% of source infections send traffic to a broker shared with other source infections. Such brokers on average send traffic to 2.4 different pharmacies.

Traffic broker characteristics (Table 4). Unsurprisingly, in light of what we saw above, 61.1% of brokers drive traffic to a single pharmacy, receiving traffic from 18.9 infected URLs on average. 33.8% of brokers redirect to multiple pharmacies, and receive on average traffic from 11.8 URLs. Finally only 5.2% of traffic brokers send traffic to other traffic brokers.

Pharmacies (Table 5). We see that 56% of pharmacies do not rely on any broker and get their traffic, on average, from 4.6 infected URLs. 17.8% of all pharmacies get traffic from a dedicated broker, which feeds them traffic coming from about 24.2 distinct infected URLs. Slightly less than a third of all pharmacies use a shared traffic broker, which—interestingly enough—forward traffic from only 5.2 infected URLs. In other words, dedicated traffic brokers appear to be driving considerably more traffic than "co-hosted" solutions using shared traffic brokers. This in turn seems to give further credence to the belief that "advertising networks" (e.g., pharmaceutical affiliates) are highly heterogeneous, with actors ranging from powerful "dedicated" brokers to others operating on a shoe-string budget. The proportion of pharmacies directly linked to infections, without a traffic broker, is high – and can be explained by the difficulties search-redirection attacks experienced in 2013, and evidenced in Figure 2.

Network characteristics. Table 6 provides an overview of the graphs consisting of all redirection chains on any given day. We observe a very strong network heterogeneity, with large connected components that appear to dominate the graph. In other words, the illicit advertising business is dominated by a few large players. The same observation was reported in earlier work [15, 21].

It is worth examining whether this concentration in advertisers changes over time. Figure 10 provides some elements of answer. We plot, as a function of time the maximum (top) and average (bottom) degree of traffic brokers and destinations. The degree is defined here as the sum of the number of links going in (in-degree) and out (out-degree) of a given "node" (traffic broker or destination). Each datapoint represents a 7-day moving average. The vertical lines correspond to the events introduced in Section 5. The size of the largest traffic brokers varies drastically over time—the spikes observed in late 2012 seem to have been caused by particularly virulent campaigns (where a few brokers received a large amount of traffic from many infected sites) that took time to be fended off by search engines. Since early 2013, the size of the largest brokers has decreased a fair bit, reflecting the trend that search-redirection might be less popular than it was in 2012.

Shared infrastructure. We complete our analysis of the redirection network by looking at the traffic brokers used for different (non-pharmaceutical) types of trades, and the extent to which they overlap with the pharmaceutical trade. Table 7 gives an overview of these results over the time interval 10/31/2011–09/16/2013. Over a long enough time interval, there is modest overlap between the various types of products. Source infections are rarely used for multiple campaigns; traffic broker domains tend to show a bit more overlap, presumably due to the fact that miscreants take advantage of lax verification policies at certain hosting providers. At the FQDN level, though, both destinations (i.e., shops) and brokers show little evidence of overlap, which is surprising given the known fact that certain botnets operate over multiple markets. Even in such cases, the different business domains appear to be kept separate.

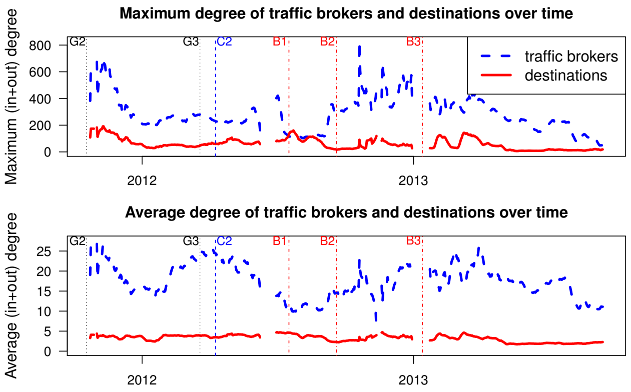


Figure 10: Maximum and average degree of traffic brokers and destinations over time.

Type and granularity of node	Drugs	Other mkts. combined	Shared #	Jaccard index (%)
Source infection FQDNs	14 770	3 975	167	0.9
Traffic broker Domains	382	202	34	6.2
Traffic broker FQDNs	735	297	33	3.3
Destination domains	2 232	1 388	120	3.4
Destination FQDNs	2 249	1 388	119	3.4

Table 7: Overlap in the criminal infrastructures. The fourth column is computed as the Jaccard index between the two sets.

8. LIMITATIONS

In addition to the numerous difficulties one faces when dealing with such long-range datasets, this study presents two major limitations. First, we have only looked at Google results. We justify this by the market share dominance of Google, at least in the US [4], but point out that other studies [2, 22] have shown other search engines are not immune to search-poisoning. Second, we have mostly looked at search results based on their presence or not in the result corpora. What is more important, however, is their *position* in the results. While top links are frequently clicked on, it has been shown that links past the tenth result have close to zero probability of being used [9]. Weighing the results we obtained by click probability would probably yield a better insight into which operations are profitable. We do note, however, that in our previous work we show that the *type* of results (e.g., search-redirection attacks vs. health resources) remained fairly consistent regardless of position [15]. Our brief examination of top-10 results in periods T_2 and T_3 confirms that active redirects and direct links to unlicensed pharmacies appear frequently among top results, and are thus expected to drive significant amounts of traffic.

9. CONCLUSIONS

Search engines are invaluable tools that deliver enormous value to consumers by referring them to the most relevant resources quickly and effortlessly. Search-engine poisoning threatens to undermine this value proposition, and could conceivably lead users to reduce their online activities [1].

We have presented the results of a longitudinal, large-scale empirical investigation into search-engine poisoning. Our long-term view has enabled us to draw several new and important insights. First, despite the best efforts of search engines to demote low-quality content and browsers to protect the privacy of search queries, miscreants have readily adapted. In fact, the share of results taken over by search-redirection attacks doubled from late 2010 to late 2012, before falling slightly. Second, efforts to clean up the compromised websites that initiate the redirections have improved: the persistence of source infections has steadily fallen from one month to two weeks. But here too, the attackers have adapted, notably by simply compromising more websites. Third, we continue to observe extensive concentration in the funneling of traffic from source infections to destinations via a small number of central brokers.

A key takeaway from this investigation is that uncoordinated interventions by individual stakeholders – a search engine ranking algorithm tweak here, a push by some hosting providers to clean up infected servers there – is not sufficient to disrupt persistent poisoning attempts. Examining this problem using Crime Script Analysis [5], Leontiadis has shown that focusing instead on key points of concentration and in cooperation across stakeholders is required to have measurable impact [14]. For instance, coordinated traffic broker take-downs at the AS level, in conjunction with the demotion or removal of poisoned search results at the search engine level (e.g., using proactive identification techniques [25]), could impact the economics of search engine poisoning significantly, and, hopefully, durably.

10. ACKNOWLEDGMENTS

We thank our anonymous reviewers and (the non-anonymous) Stuart Schechter for feedback on earlier revisions of this manuscript. This research was partially supported by the National Science Foundation under ITR award CCF-0424422 (TRUST) and SaTC award CNS-1223762; and by the Department of Homeland Security Science and Technology Directorate, Cyber Security Division (DHS S&T/CSD), the Government of Australia and SPAWAR Systems Center Pacific via contract number N66001-13-C-0131. This paper represents the position of the authors and not that of the aforementioned agencies.

11. REFERENCES

[1] R. Anderson, C. Barton, R. Böhme, R. Clayton, M. van Eeten, M. Levi, T. Moore, and S. Savage. Measuring the cost of cybercrime. In *Proc. (online) WEIS 2012*. Berlin, Germany, June 2012.

[2] K. Borgolte, C. Kruegel, and G. Vigna. Delta: automatic identification of unknown web-based infection campaigns. In *Proc. ACM CCS 2013*, pages 109–120, Berlin, Germany, November 2013.

[3] T. Catan. Google forks over settlement on Rx ads. *The Wall Street Journal*, August 2011. Available online at `http://online.wsj.com/news/articles/SB10001424053111904787404576528332418595052`.

[4] comScore. February 2014 US search engine rankings. `https://www.comscore.com/Insights/Press_Releases/2014/3/comScore_Releases_February_2014_U.S._Search_Engine_Rankings`, 2014. Last accessed August 26, 2014.

[5] D. Cornish. The procedural analysis of offending and its relevance for situational prevention. *Crime prevention studies*, 3:151–196, 1994.

[6] R. Dingledine, N. Mathewson, and P. Syverson. Tor: The second-generation onion router. In *Proc. USENIX Security 2004*. San Diego, CA, August 2004.

[7] Z. Gyöngyi and H. Garcia-Molina. Link spam alliances. In *Proc. ACM VLDB 2005*, pages 517–528, Trondheim, Norway, August 2005.

[8] AOL Inc. Open Directory project. `http://www.dmoz.org/`.

[9] T. Joachims, L. Granka, B. Pan, H. Hembrooke, and G. Gay. Accurately interpreting clickthrough data as implicit feedback. In *Proc. ACM SIGIR '05*, pages 154–161, Salvador, Brazil, 2005.

[10] J. John, F. Yu, Y. Xie, M. Abadi, and A. Krishnamurthy. deSEO: Combating search-result poisoning. In *Proc. USENIX Security 2011*, San Francisco, CA, August 2011.

[11] C. Kanich, C. Kreibich, K. Levchenko, B. Enright, G. Voelker, V. Paxson, and S. Savage. Spamalytics: An empirical analysis of spam marketing conversion. In *Proc. ACM CCS 2008*, Alexandria, VA, October 2008.

[12] E. Kao. Making search more secure, October 2011. `http://googleblog.blogspot.com/2011/10/making-search-more-secure.html`.

[13] E. Kaplan and P. Meier. Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, 53:457–481, 1958.

[14] N. Leontiadis. *Structuring disincentives for online criminals*. PhD thesis, Carnegie Mellon University, 2014.

[15] N. Leontiadis, T. Moore, and N. Christin. Measuring and analyzing search-redirection attacks in the illicit online prescription drug trade. In *Proc. USENIX Security 2011*, San Francisco, CA, August 2011.

[16] N. Leontiadis, T. Moore, and N. Christin. Pick your poison: Pricing and inventories at unlicensed online pharmacies. In *Proc. ACM EC*, pages 621–638, Philadelphia, PA, June 2013.

[17] K. Levchenko, N. Chachra, B. Enright, M. Felegyhazi, C. Grier, T. Halvorson, C. Kanich, C. Kreibich, H. Liu, D. McCoy, A. Pitsillidis, N. Weaver, V. Paxson, G. Voelker, and S. Savage. Click trajectories: End-to-end analysis of the spam value chain. In *Proc. 2011 IEEE Symposium on Security and Privacy*, Oakland, CA, May 2011.

[18] Z. Li, S. Alrwais, X. Wang, and E. Alowaisheq. Hunting the red fox online: Understanding and detection of mass redirect-script injections. In *Proc. 2014 IEEE Symposium on Security and Privacy*, San Jose, CA, May 2014.

[19] Legitscript LLC. Legitscript pharmacy validation. `http://www.legitscript.com/pharmacies/`.

[20] L. Lu, R. Perdisci, and W. Lee. SURF: Detecting and measuring search poisoning. In *Proc. ACM CCS 2011*, Chicago, IL, October 2011.

[21] D. McCoy, A. Pitsillidis, G. Jordan, N. Weaver, C. Kreibich, B. Krebs, G. Voelker, S. Savage, and K. Levchenko. Pharmaleaks: Understanding the business of online pharmaceutical affiliate programs. In *Proc. USENIX Security 2012*, Bellevue, WA, August 2012.

[22] T. Moore, N. Leontiadis, and N. Christin. Fashion crimes: Trending-term exploitation on the web. In *Proc. ACM CCS 2011*, Chicago, IL, October 2011.

[23] J. Mueller. Upcoming changes in Google's HTTP referrer, March 2012. `http://googlewebmastercentral.blogspot.com/2012/03/upcoming-changes-in-googles-http.html`.

[24] A. Singhal and M. Cutts. Finding more high-quality sites in search, February 2011. `http://googleblog.blogspot.com/2011/02/finding-more-high-quality-sites-in.html`.

[25] K. Soska and N. Christin. Automatically detecting vulnerable websites before they turn malicious. In *Proc. USENIX Security 2014*, San Diego, CA, August 2014.

[26] The Internet Archive. Wayback machine. `https://archive.org/web/`.

[27] VirusTotal. Free Online Virus, Malware and URL Scanner. `https://www.virustotal.com/`.

[28] D. Wang, M. Der, M. Karami, L. Saul, D. McCoy, S. Savage, and G. Voelker. Search + seizure: The effectiveness of interventions on seo campaigns. In *Proc. ACM IMC'14*, Vancouver, BC, Canada, November 2014.

[29] D. Wang, G. Voelker, and S. Savage. Juice: A longitudinal study of an SEO botnet. In *Proc. NDSS'13*, San Diego, CA, February 2013.